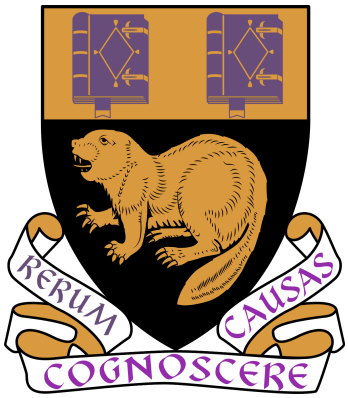
Spatial Modelling and Volatility Matrix Estimation in High Dimension Statistics with Financial Applications



Cheng Qian

The Department of Statistics

London School of Economics and Political Science

A thesis submitted for the degree of

Doctor of Philosophy

December 2018

This thesis is dedicated to
Ruijia Zhan

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I confirm that Chapter 1 and 3 were jointly co-authored with my supervisor, Dr.Clifford Lam and Chapter 2 was jointly co-authored with Dr.Clifford Lam and Professor Hui Wang from Central University of Finance and Economics, China.

Acknowledgements

First of all, I would like to thank my supervisor, Dr.Clifford Lam, for this constant help, patient guidance and professional advice for my PhD research and life. Without him, it is impossible that I can finish my PhD study early and pursue a postdoc in the following two years.

I also want to appreciate my advisor, Professor Qiwei Yao, who provides me plenty of opportunities during my PhD study. His cleverness and kindness impress me deeply. Besides my two supervisors, I need to show my gratitude to all faculties and staffs in our department for providing me a wonderful environment for my study. Especially, Dr.Xinghao Qiao and Dr.Yining Chen help the most and guide me to find my future.

Last but not the least, thanks my wife Ruijia Zhan and my parents Yun Qian and Chunxiang Xu for their always love and support.

Abstract

High dimension modelling is an important area in modern statistics. For example, a large number of problems that arise in finance are also inspired by more and more available high dimensional data. The main objective of this thesis is to investigate three methodologies in high dimension statistics with the application in finance. The subsequent chapters are organized as follow. The first two chapters are about spatial modeling and its inference respectively. The third chapter tackles a different problem about the estimation of large integrated volatility matrix of high frequency data.

In the first chapter, a dynamic spatial model with different weight matrices for different time-lagged spatial effects is proposed. Unlike assuming a known spatial weight matrix, the proposed method estimates each spatial weight matrix for corresponding spatial effect by a linear combination of a set of specified spatial weight matrices to avoid misspecification. To estimate the coefficients for linear combinations and covariates, the profiled least square estimation is used with instrumental-like variables. A further selection on spatial weight matrices is introduced by adding an adaptive LASSO penalty on the coefficients of linear combination. All theoretical results are built on the scenario when the sample size T and panel dimension N go to infinity. The functional dependence in time series proposed by Wu (2005) is applied for the asymptotic normality of the estimated parameters. The oracle properties for model selection are developed including the asymptotic normality and sign consistency. Apart from a simulated data used to illustrate the performance of the proposed model, we also apply the proposed model to 32 important stocks from the Euro Stoxx 50 and S&P 500 in 2015 to invest the spatial interaction of them.

The second chapter discusses the inference for the spatial dynamic model.To estimate the spatial weight matrices for contemporaneous and time-lagged spatial effects, two linear combinations of a set of the specified spatial weight matrices are adopted respectively. We extend the quasi-maximum likelihood estimation for the linear combination coefficients

in our model and their consistency and asymptotic normality are established when both N and T are large. Using the asymptotic normality of the quasi-maximum likelihood estimators, a Wald test can be employed on the coefficients of the linear combination. Then, a diagnostic test proposed in Chang et al. (2017) is applied to test whether the fitted residuals perform like a white noise vector in our large N and large T setting. Simulated and real data are used to demonstrate the performance of the proposed quasi-maximum likelihood estimation and all above tests.

The third chapter is about the estimation of large integrated volatility matrix for high frequency data. Besides the microstructure noises and non-synchronous trading times for high frequency data analysis should be fixed, the bias in the extreme eigenvalues coming from the high dimensionality are also not negligible. A nonparametric eigenvalue regularization proposed in Lam (2016) is applied on three existing volatility matrix estimators, such as multi-scale, kernel and pre-averaging realized volatility matrix estimators. One advantage for the proposed estimators is no need for implicit assumptions on the structure of the true integrated volatility matrix. It can be proved that the bias in the extreme eigenvalues can be shrunk and the regularized volatility estimators are positive definite in probability. Incidentally, the bias-corrected versions of kernel and pre-averaging estimators, which have faster rate of convergence at $n^{-1/4}$ but are not guaranteed to be positive definite in Barndorff-Nielsen et al. (2011) and Christensen et al. (2010) respectively, are now regularized to be positive definite in probability, and we prove their rates of convergence to an "ideal" estimator under the spectral norm are also at $n^{-1/4}$ under $p/n \to c > 0$. Jump and its removal by wavelet method in Fan and Wang (2007) are also included and all theoretical results are still hold. All proposed methods are applied on the simulated data. We also test the performance of the proposed methods on the stocks from the list "Fifty Most Active Stocks on NYSE" and "Fifty Most Active Stocks by Dollar Volume on NYSE".

Contents

List of Figures

Chapter 1

Spatial Lag Model with Time-lagged Effects and spatial weight Matrix Estimation

1.1 Introduction

There are always complicated correlation over cross section and time in the real data. In econometrics, spatial econometrics develop the models to investigate the cross-sectional interactions. For example, spatial autoregressive model proposed in Cliff and Ord (1973) is very powerful. Among these models, Anselin et al. (2008) divides them into four groups. The first type is "pure space recursive" if only a spatial time lag is included. The second type is "time-space recursive" if both an individual time lag and a spatial time lag are included. The third type is "time-space simultaneous" if an individual time lag and a contemporaneous spatial lag are specified. And finally, the last type is "time-space dynamic" if all forms of lags are included. Besides the spatial autoregressive model, spatial disturbance autoregressive model is considered in Elhorst (2005). All these models have been frequently used well in many fields like regional markets in Keller and Shiue (2007), labour economics in Foote (2007) or public economics in Franzese and Hays (2007), to name but a few areas.

It is obvious that spatial autoregressive model becomes one of the most active fields in econometrics and also receives considerable attention from a number of other fields. However, the correct usage of spatial autoregressive model depends on the

correct choice of spatial weight matrix whose elements reflect the strength of interaction among units in the panel. In plenty of applications of spatial autoregressive model, the spatial weight matrix is assumed as a prior knowledge. Physical distance between two units in the panel is often regarded as the inverse measurement of the interaction between them, so people use d^{-1} as the entry of the spatial weight matrix where d is the corresponding distance. Naturally, physical distance is not the only choice for the cross sectional interaction measurement. For example, in economics, two countries from a same economic organization may have a strong relation in spite of far distance between them. Even if only considering the physical distance, d^{-2} or d^{-3} can also be the candidates. Therefore, it is limited to only use one spatial weight matrix into the model. Corrado and Fingleton (2012) also criticizes the spatial econometrics due to this misspecification of spatial weight matrix. In Lam and Souza (2015b), an error upper bound is given for the estimation of spatial regression parameters, which shows that misspecification of the spatial weight matrix can introduce large bias in the final estimates.

To avoid the misspecification of the spatial weight matrix in spatial model, nonparametric models are applied in past researches, see Tran and Yakowitz (1993) and Hallin et al. (2004) for instance. The Nadaraya-Watson kernel estimator is frequently used for nonparametric regression in econometrics. For example, Robinson (2011) establishes its consistency and asymptotic distribution theory in a framework designed for various kinds of spatial data. Koroglu and Sun (2016a) improves the estimation accuracy by applying a nonparametric two-stage least squares estimation. More specifically, the second-step estimators of the unknown functional coefficients are estimated by local linear regression. However, Kostov (2013) shows that it can lead to reduced efficiency of the estimators when the sample size is small. On the other hand, there are also some works assuming the structure of spatial weight matrix. For example, Bhattacharjee and Jensen-Butler (2013) proposes to estimate such a spatial weight matrix with a symmetric assumption, while Lam and Souza (2016a) proposes to estimate the block pattern in such a matrix. With the development of high dimensional statistics, Ahrens and Bhattacharjee (2015a) considers a two-step LASSO estimation, which is based on the sparsity assumption of the spatial weight matrix. Meanwhile, adaptive LASSO is used in Lam and Souza (2015a) to estimate the spatial weight matrix together with fixed effects in the spatial model.

All methods mentioned above only include contemporaneous spatial effect but no time-lagged spatial effect to avoid the complexity of the modelling, as one more time-lagged spatial effect means one more spatial weight matrix should be estimated. However, from the evidence in our real data example, including time-lagged spatial effect is necessary. In our proposed dynamic spatial model, each spatial

weight matrix involved is estimated by a linear combination of user-specified spatial weight matrices. A similar model in Lee and Liu (2010b) is named as high order spatial autoregressive model. The inclusion of more than one spatial weight matrices can allow spatial dependence from different interaction characteristics such as geographical contignity and economic interaction. Therefore, it helps to avoid the risk of misspecification of the spatial weight matrices, while maintaining the overall parsimony of the model.

As for the high order spatial autoregressive model, Lee and Liu (2010b) and Lee and Yu (2014a) propose the generalized method of moments estimation. However, as discussed in Li (2017), when we have a large or moderately large sample size T, the generalized method of moments gets into trouble of "many moments bias" as the number of moment conditions also increases dramatically. This point is also mentioned in Lee and Yu (2014a) that generalized method of moments requires careful analysis when $T \to \infty$. Another solution for the high order spatial autoregressive model is quasi-maximum likelihood estimation introduced in Yu et al. (2008) and Li (2017). But it is well known that quasi-maximum likelihood estimation is not practical and computationally infeasible because of the complex parameter space and the difficulty in Jacobian determinant evaluation, especially for the proposed model in this chapter that includes p time-lagged spatial effects.

Therefore, it is prefered to find an efficient estimation but no complicated parameter space is involved. Therefore, the profiled least square estimation is applied for the coefficients of linear combinations and covariates. Meanwhile, the innate endogeneity in our time-lagged spatial model causing least square type estimation inconsistent can not be ignored. To overcome this difficulty, we introduce instrument-like variables. In the particular case when the covariates are exogenous, they themselves can act as these instrument-like variables. We estimate the "best" linear combination for each required spatial weight matrix, then an adaptive LASSO can be applied for highlighting the relative contributions of each specified one. The convergence and asymptotic normality of all estimators are presented under the functional dependence measurement of time series variables in Wu (2005) or Wu (2011), allowing both the sample size T and the panel size N to grow to infinity together. As shown in our real data analysis, with the input of different specified spatial weight matrices, the scope of applications of our model is expanded since there are numerous ways to specify a spatial weight matrix.

The rest of the chapter is organized as follows. Section 1.2 introduces our methodology, including the model and the estimation method. Properties of our estimators, including asymptotic normality are presented in Section 1.3. Simulation results and real data analysis are reported, respectively, in Section 1.5 and 1.6. The conclusion

and some relative further studies are listed in Section 1.7. All the technical proofs are relegated to the Section 1.8.

1.2 Methodology

1.2.1 The Model

Consider the following dynamic spatial lag model

$$y_t = \boldsymbol{\mu} + \boldsymbol{W}_0 y_t + \boldsymbol{W}_1 y_{t-1} + \cdots + \boldsymbol{W}_p y_{t-p} + \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T, \tag{1.1}$$

where $y_t = (y_{t1}, y_{t2}, ..., y_{tN})^T$ is an $N \times 1$ vector of observed time series variables and $\boldsymbol{\mu}$ is an $N \times 1$ constant vector. The data starts from y_{1-p}, and hence the true sample size is $T + p$. It does not affect our asymptotic analysis since p is finite in this paper. Hereafter when we talk about the sample size, we use T instead of $T + p$ for simplicity. For $j = 0, 1, \ldots, p$, $\boldsymbol{W}_j$ is an $N \times N$ spatial weight matrix with 0 on the main diagonal, which model the simultaneous and dynamic interaction between different unit in the panel. To capture the dynamic interaction between same unit, the $N \times K$ matrix of covariates $\boldsymbol{X}_t$ can contain y_{t-j} for $j = 1, \ldots, p$ in its columns on top of other covariates, while $\boldsymbol{\beta}$ is the $K \times 1$ vector of regression coefficients. The series $\{\boldsymbol{\epsilon}_t\}$ is an innovation process with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_\epsilon$. For more detailed assumptions, see Section 1.8.1.

In many applied spatial econometrics applications, $\boldsymbol{W}_0$ is assumed known and there are no lagged terms $\boldsymbol{W}_j y_{t-j}$ for $j = 1, \ldots, p$. Instead of assuming all the spatial weight matrices are known, in this paper we assume that there are M specified spatial weight matrices $\boldsymbol{W}_{0i}$, $i = 1, \ldots, M$, such that each spatial weight matrix is a linear combination of the M specified ones. This is motivated by the fact that there are often more than one measures of spatial interactions. For instance, for the geographical distance r alone between two specific locations, we can specify three different entries r^{-1}, r^{-2} and r^{-3}, creating three specified spatial weight matrices. These are indeed our distance specifications included in our data application in Section 1.6. Spatial contiguity is also another popular choice in spatial econometrics. The linear combination for each $\boldsymbol{W}_j$ is written as

$$\boldsymbol{W}_j = \sum_{i=1}^{M} \delta_{ji} \boldsymbol{W}_{0i},$$

where δ_{ji} for $i = 1, \ldots, M,\ j = 0, \ldots, p$ are unknown coefficients in the proposed model.

Apart from allowing for estimating the spatial weight matrices from pre-specified ones, our model also includes time-lagged spatial effects. In a differently specified spatial lag model, Dou et al. (2016) includes one lag to reflect such effects. We generalize this to p time-lagged effects, with p to be determined by data driven methods as described in Section 1.4. The pure dynamic effects are captured by the term $\boldsymbol{X}_t\boldsymbol{\beta}$, since we can allocate $\{y_{t-1}, \ldots, y_{t-p}\}$ to be the columns in $\boldsymbol{X}_t$, so that then $K \geq p$, and $K = p$ if no other covariates are present. Not counting the parameters in $\boldsymbol{\mu}$, there are $K + M(p+1)$ parameters to be estimated in total.

With $\boldsymbol{\mu}$, the spatial fixed effects of the model is then $(\boldsymbol{I}_N - \boldsymbol{W}_0)^{-1}\boldsymbol{\mu}$. For identifiability of such, we assume without loss of generality that $\mathbb{E}(X_t) = 0$. As the instrumental variable deducts its mean in our methodology, the non-zero μ can be removed. Therefore, our assumption $\mathbb{E}(X_t) = 0$ is same as $\mathbb{E}(y_t) = 0$. If we do not have $\mathbb{E}(X_t) = 0$, we can write

$$\boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\mu} = (\boldsymbol{X}_t - \mathbb{E}(\boldsymbol{X}_t))\boldsymbol{\beta} + (\boldsymbol{\mu} + \mathbb{E}(\boldsymbol{X}_t)\boldsymbol{\beta})$$

so that the spatial fixed effects are now captured by $\boldsymbol{\mu} + \mathbb{E}(\boldsymbol{X}_t)\boldsymbol{\beta}$ rather than $\boldsymbol{\mu}$.

1.2.2 Profiled least square estimation with endogeneity

The first important problem needed to be concerned is the endogeneity in model (1.1). To estimate $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ more efficiently, we assume that there are variables $\boldsymbol{B}_t$ of size $N \times K$ such that they are correlated with $\boldsymbol{X}_t$ but independent of $\boldsymbol{\epsilon}_t$ for each $t = 1, \ldots, T$. In particular, if $\boldsymbol{X}_t$ is exogenous, we can set $\boldsymbol{B}_t = \boldsymbol{X}_t$. To apply the instrument-like variable $\boldsymbol{B}_t$ in the vectorized model, we first define

$$\boldsymbol{B} = T^{-1/2}N^{-a/2}(\boldsymbol{B}_{\boldsymbol{\zeta}} - \overline{\boldsymbol{B}}_{\boldsymbol{\zeta}}) = T^{-1/2}N^{-a/2}\boldsymbol{I}_N \otimes \{(\boldsymbol{I}_T \otimes \boldsymbol{\zeta}^T)(\boldsymbol{B}_1 - \bar{\boldsymbol{B}}, \ldots, \boldsymbol{B}_T - \bar{\boldsymbol{B}})^T\},$$

where $\bar{\boldsymbol{B}} = T^{-1}\sum_{t=1}^{T}\boldsymbol{B}_t$, $\boldsymbol{\zeta} = K^{-1}1_K$ and 1_K is $K \times 1$ vector of ones. We set $a = 1$ in our algorithms. In theory, it is there only to adjust the order of eigenvalues of some constructs involving $\boldsymbol{B}$. See the technical assumptions in Section 1.8 for more details. The value of $\boldsymbol{\zeta}$ is also not the only choice, and we will introduce a way to choose a data driven one in Section 1.4.3.

Using $\boldsymbol{B}$ to avoid the inconsistency from endogeneity, we first rewrite (1.1) to present our model more neatly as

$$y = \boldsymbol{\mu} \otimes 1_T + \boldsymbol{Z}_0\boldsymbol{V}_0\boldsymbol{\delta}_0 + \boldsymbol{Z}_1\boldsymbol{V}_0\boldsymbol{\delta}_1 + \cdots + \boldsymbol{Z}_p\boldsymbol{V}_0\boldsymbol{\delta}_p + \boldsymbol{X}_{\boldsymbol{\beta}}\text{vec}(\boldsymbol{I}_N) + \boldsymbol{\epsilon},$$

where $y = \text{vec}(y_1, \ldots, y_T)^T$, $\boldsymbol{\epsilon} = \text{vec}(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T)^T$, $\boldsymbol{Z}_j = \boldsymbol{I}_N \otimes (y_{1-j}, \ldots, y_{T-j})^T$ and $\boldsymbol{\delta}_j = (\delta_{j1}, \delta_{j2}, \ldots, \delta_{jM})^T$ for $j = 0, 1, \ldots, p$, $\boldsymbol{X}_{\boldsymbol{\beta}} = \boldsymbol{I}_N \otimes (\boldsymbol{I}_T \otimes \boldsymbol{\beta}^T)(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)^T$, and $\boldsymbol{V}_0 = (\text{vec}(\boldsymbol{W}_{01}^T), \ldots, \text{vec}(\boldsymbol{W}_{0M}^T))$. The notation $\otimes$ is the Kronecker product, and 1_T defines a vector of ones with size T. Simplifying, we have

$$y = \boldsymbol{\mu} \otimes 1_T + \boldsymbol{Z}\boldsymbol{V}\boldsymbol{\delta} + \boldsymbol{X}_{\boldsymbol{\beta}}\text{vec}(\boldsymbol{I}_N) + \boldsymbol{\epsilon}, \tag{1.2}$$

where $\boldsymbol{Z} = (\boldsymbol{Z}_0, \ldots, \boldsymbol{Z}_p)$, $\boldsymbol{\delta} = (\boldsymbol{\delta}_0^T, \ldots, \boldsymbol{\delta}_p^T)^T$, and $\boldsymbol{V} = \boldsymbol{I}_{p+1} \otimes \boldsymbol{V}_0$. Then, multiplying $\boldsymbol{B}^T$ on both sides of (1.2), we arrive at the augmented model

$$\boldsymbol{B}^T y = \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}\boldsymbol{\delta} + \boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}}\text{vec}(\boldsymbol{I}_N) + \boldsymbol{B}^T\boldsymbol{\epsilon}. \tag{1.3}$$

The constant term disappears since $\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{\mu} \otimes 1_T) = \boldsymbol{0}$. Removing the N-dimensional constant term makes estimation much easier, while the error term $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\epsilon}$ is now weaker in correlations with the design matrix $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{V}$, so that least square estimation becomes viable again.

After introducing $\boldsymbol{B}$ serving as instrumental variable, same as Lam and Souza (2018), we can apply profiled least square estimation on the augmented model to avoid the nonlinearity and to reduce variance if we estimate $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ simultaneously. More specifically, to profile out $\boldsymbol{\beta}$ and estimate $\boldsymbol{\delta}$, we rewrite the augmented model as

$$\boldsymbol{B}^{vT}y_0^v = \boldsymbol{B}^{vT}(\sum_{i=1}^{M}\delta_{0i}\boldsymbol{W}_{0i}^{\otimes})y_0^v + \boldsymbol{B}^{vT}\sum_{j=1}^{p}(\sum_{i=1}^{M}\delta_{ji}\boldsymbol{W}_{0i}^{\otimes})y_j^v + \boldsymbol{B}^{vT}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}^{vT}\boldsymbol{\epsilon}^v,$$

where $y_j^v = (y_{1-j}^T, \ldots, y_{T-j}^T)^T$ for $j = 0, 1, \ldots, p$, $\boldsymbol{\epsilon}^v = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_T^T)^T$, $\boldsymbol{B}^v = ((\boldsymbol{B}_1 - \bar{\boldsymbol{B}})^T, \ldots, (\boldsymbol{B}_T - \bar{\boldsymbol{B}})^T)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_T^T)^T$ and $\boldsymbol{W}_{0i}^{\otimes} = \boldsymbol{I}_T \otimes \boldsymbol{W}_{0i}$ for $i = 1, \ldots, M$. Assuming $\boldsymbol{\delta}$ is known, we can estimate $\boldsymbol{\beta}$ by the least squared method, resulting in

$$\boldsymbol{\beta}(\boldsymbol{\delta}) = (\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\Big\{(\boldsymbol{I}_{TN} - \sum_{i=1}^{M}\delta_{0i}\boldsymbol{W}_{0i}^{\otimes})y_0^v - \sum_{j=1}^{p}(\sum_{i=1}^{M}\delta_{ji}\boldsymbol{W}_{0i}^{\otimes})y_j^v\Big\}. \tag{1.4}$$

This formula provides a basis for a profile least square estimator for $\boldsymbol{\delta}$. We can show that by substituting the above into the augmented model (1.3), the profile least square estimator for $\boldsymbol{\delta}$ is

$$\hat{\boldsymbol{\delta}} = \{(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\}^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{K}y_0^v - \boldsymbol{B}^T y), \tag{1.5}$$

where

$$\boldsymbol{K} = T^{-1/2}N^{-a/2}(\sum_{t=1}^{T}\boldsymbol{X}_t \otimes (\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\zeta})(\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT},$$

$$\boldsymbol{H} = \boldsymbol{K}\left[\boldsymbol{W}_{01}^{\otimes}, \ldots, \boldsymbol{W}_{0M}^{\otimes}\right](\boldsymbol{I}_M \otimes y_0^v, \boldsymbol{I}_M \otimes y_1^v, \ldots, \boldsymbol{I}_M \otimes y_p^v).$$

Therefore, with $\hat{\boldsymbol{\delta}}$, the profile least square estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{\boldsymbol{\delta}}) = (\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\Big\{(\boldsymbol{I}_{TN} - \sum_{i=1}^{M}\hat{\delta}_{0i}\boldsymbol{W}_i^{\otimes})y_0^v - \sum_{j=1}^{p}(\sum_{i=1}^{M}\hat{\delta}_{ji}\boldsymbol{W}_i^{\otimes})y_j^v\Big\}. \tag{1.6}$$

Finally, to estimate $\boldsymbol{\mu}$, we can use

$$\hat{\boldsymbol{\mu}} = \Big(\boldsymbol{I}_N - \sum_{j=0}^{p}\widehat{\boldsymbol{W}}_j\Big)\bar{y} - \bar{X}\hat{\boldsymbol{\beta}}, \quad \text{where } \widehat{\boldsymbol{W}}_j = \sum_{i=1}^{M}\hat{\delta}_{ji}\boldsymbol{W}_{0i}.$$

The corresponding spatial fixed effects estimator is then given by $(\boldsymbol{I}_N - \widehat{\boldsymbol{W}}_0)^{-1}\hat{\boldsymbol{\mu}}$.

1.2.3 Selection of specified spatial weight matrices

As the specified spatial weight matrices $\boldsymbol{W}_{0i}$ are arbitrary, it is not necessary that all of them should be included. Since $\widehat{\boldsymbol{\delta}}$ is a least square-type estimator, each element in it is not estimated to be exactly 0 in general. This hinders the selection of the specified spatial weight matrices, which is important for us to see which one truely contributes to the overall spatial weight matrix and which one does not. Especially, the different time-lagged spatial effects shown by the corresponding spatial weight matrix $\boldsymbol{W}_j$ can have the different selection on the specified spatial weight matrices. It is true that some spatial characteristics can have the delayed impact, which is also reflected by our real data example in Section 1.6.

As a classical model selection method, LASSO needs irrepresentable condition to make the estimator sparse and have sign consistency. As discussed in Lam and

Souza (2018), the tuning parameter used in LASSO is same for all element, which causing excessive penalization by larger tuning parameter or insufficient penalization by smaller tuning parameter. To resolve this, Zou (2006) propose a method named adaptive LASSO. On the other hand, compared with SCAD penalty proposed in Fan and Li (2001), adaptive LASSO still enjoys convexity. Therefore, it has computational efficiency and can apply the same algorithm as standard LASSO.

To handle the selection of specified spatial weight matrices, we apply the penalized profiled least square estimator $\widetilde{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$ same as the way used in Lam and Souza (2018), with

$$\tilde{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \frac{1}{2T} \|\boldsymbol{B}^T y - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V} \boldsymbol{\delta} - \boldsymbol{B}^T \boldsymbol{X}_{\boldsymbol{\beta}(\boldsymbol{\delta})} \operatorname{vec}(\boldsymbol{I}_N)\|^2 + \gamma_T \boldsymbol{u}^T |\boldsymbol{\delta}|, \tag{1.7}$$

where $\boldsymbol{u} = (|\widehat{\delta}_{0,1}|^{-1}, \ldots, |\widehat{\delta}_{0,M}|, \ldots, |\widehat{\delta}_{p,1}|^{-1}, \ldots, |\widehat{\delta}_{p,M}|^{-1})^T$, and $|\boldsymbol{\delta}|$ represents the same vector $\boldsymbol{\delta}$ with all its entries taken absolute value. The penalty term in (1.7) is similar to the adaptive LASSO proposed in Zou (2006), which shows a better performance in model selection than standard LASSO. A more direct penalized least square formulation is given by

$$\widetilde{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \frac{1}{2T} \|\boldsymbol{B}^T y - (\boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V} - \boldsymbol{H})\boldsymbol{\delta} - \boldsymbol{g}\|^2 + \gamma_T \boldsymbol{u}^T |\boldsymbol{\delta}|, \text{ where}$$

$$\boldsymbol{g} = T^{-1/2} N^{-a/2} (\sum_{t=1}^{T} \boldsymbol{X}_t \otimes (\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\zeta})(\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} y^v.$$

The tuning parameter γ_T can be found in Assumption R6 in Section 1.8. For choosing an appropriate γ_T in practice, see Section 1.4.2.

1.3 Theoretical Properties

We define some notations first before we present the time dependence measurement from Wu (2005) and our theoretical properties of all estimators. Compared with classic mixing condition, Wu (2005) discusses that, for stationary causal processes, the calculation of probabilistic dependence measures is generally not easy because it involves the complicated manipulation of taking the supremum over two sigma algebras. Additionally, many well-known processes are not strong mixing.

Let $\{\boldsymbol{b}_t\} = \{\operatorname{vec}(\boldsymbol{B}_t)\}$ and $\{\boldsymbol{x}_t\} = \{\operatorname{vec}(\boldsymbol{X}_t)\}$ be the vectorized processes for $\{\boldsymbol{B}_t\}$

and $\{\boldsymbol{X}_t\}$ respectively, both with length NK. For $t = 1, \ldots, T$, we assume that

$$\boldsymbol{x}_t = \{f_j(\mathcal{F}_t)\}_{1\leq j\leq NK},\ \boldsymbol{b}_t = \{g_j(\mathcal{G}_t)\}_{1\leq j\leq NK},\ \boldsymbol{\epsilon}_t = \{h_l(\mathcal{H}_t)\}_{1\leq l\leq N},$$

where the $f_j(\cdot)$, $g_j(\cdot)$ and $h_l(\cdot)$ are measurable functions defined on the real line, and let the shift process $\mathcal{F}_t = (..., e_{x,t-1}, e_{x,t})$, $\mathcal{G}_t = (..., e_{b,t-1}, e_{b,t})$ and $\mathcal{H}_t = (..., e_{\epsilon,t-1}, e_{\epsilon,t})$ are defined by independent and identically distributed processes $\{e_{x,t}\}$, $\{e_{b,t}\}$ and $\{e_{\epsilon,t}\}$ respectively, with $\{e_{b,t}\}$ independent of $\{e_{\epsilon,t}\}$ but correlated with $\{e_{x,t}\}$.

To build the asymptotic normality results for the estimators, we apply the functional dependence measure introduced in Wu (2005) for gauging the serial dependence of a process. For $d > 0$, define

$$\theta^x_{t,d,j} = \|x_{tj} - x'_{tj}\|_d = (\mathbb{E}|x_{tj} - x'_{tj}|^d)^{1/d},$$
$$\theta^b_{t,d,j} = \|b_{tj} - b'_{tj}\|_d = (\mathbb{E}|b_{tj} - b'_{tj}|^d)^{1/d},$$
$$\theta^\epsilon_{t,d,l} = \|\epsilon_{tl} - \epsilon'_{tl}\|_d = (\mathbb{E}|\epsilon_{tl} - \epsilon'_{tl}|^d)^{1/d},$$

where $j = 1, ..., NK$, $l = 1, ..., N$ and $x'_{tj} = f_j(\mathcal{F}'_t)$, $\mathcal{F}'_t = (..., e_{x,-1}, e'_{x,0}, e_{x,1}, ..., e_{x,t})$, with $e'_{x,0}$ independent of all other $e_{x,j}$'s. Hence x'_{tj} is a coupled version of x_{tj} with $e_{x,0}$ replaced by an independent and identically distributed copy $e'_{x,0}$. Intuitively, a large $\theta^x_{t,d,j}$ means that serial correlation is strong at least for variables at most time t apart. Finally, we have similar definitions for b'_{tj} and ϵ'_{tl}.

1.3.1 Main assumptions

We introduce some assumptions for our theorems to hold. First, we denote the L_1 norm $\|\boldsymbol{v}\|_1 = \sum_{i=1}^N |v_i|$ for a $N \times 1$ vector $\boldsymbol{v}$ whose ith element is v_i. More technical assumptions are moved to Section 1.8 to help the flow of this chapter.

M1. The elements in all $\boldsymbol{W}_i$'s can be negative and $\boldsymbol{W}_i$ itself can be asymmetric. Moreover, defining $S = \{s = 1, \ldots, K|\text{The } s\text{th column of } \boldsymbol{X}_t \text{ contains } y_{t-l},\ l = 1, \ldots, p\}$, we assume $\sum_{i=1}^M |\delta_{0i}| < 1$ and $\sum_{j=1}^p \sum_{i=1}^M |\delta_{ji}| + \sum_{s\in S} |\beta_s| < 1$.

M2. The processes $\{\boldsymbol{B}_t\}$, $\{\boldsymbol{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ are second-order stationary, with $\{\boldsymbol{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ having zero means, and $\{\boldsymbol{B}_t\}$ independent of $\{\boldsymbol{\epsilon}_t\}$. The tail condition $P(|Z| > v) \leq D_1 \exp(-D_2 v^q)$ is satisfied for the variables $B_{t,jk}$, $X_{t,jk}$ and $\epsilon_{t,j}$ by the same constants D_1, D_2 and q.

M3. Define

$$\Theta^x_{m,a} = \sum_{t=m}^{\infty} \max_{1\le j\le NK} \theta^x_{t,a,j},\ \Theta^b_{m,a} = \sum_{t=m}^{\infty} \max_{1\le j\le NK} \theta^b_{t,a,j},\ \Theta^\epsilon_{m,a} = \sum_{t=m}^{\infty} \max_{1\le j\le N} \theta^\epsilon_{t,a,j}.$$

Then we assume that for some $w > 2$, $\Theta^x_{m,2w}, \Theta^b_{m,2w}, \Theta^\epsilon_{m,2w} \le Cm^{-\alpha}$ with $\alpha, C > 0$ being constants that can depend on w.

M4. (Identification condition) Assume that the two sets of parameters $(\boldsymbol{\delta}^*, \boldsymbol{\beta}^*)$ and $(\boldsymbol{\delta}^o, \boldsymbol{\beta}^o)$ both satisfy the proposed model (1.2). Write $\boldsymbol{\delta} = (\delta_\ell)_{1\le\ell\le M(p+1)}$, and define the set H to be

$$H = \{\ell : \delta^*_\ell \neq 0 \text{ or } \delta^o_\ell \neq 0\}.$$

Then the identification condition is that the matrix $\boldsymbol{O}^T\boldsymbol{O}$ has all its eigenvalues uniformly bounded away from 0, where

$$\boldsymbol{O} = (T^{-1/2}\mathbb{E}(\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H), T^{-1/2}\mathbb{E}(\boldsymbol{B}^T\tilde{\boldsymbol{X}})), \text{ and}$$

$$\tilde{\boldsymbol{X}} = (\boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{T,1}, \ldots, \boldsymbol{x}_{1,N}, \ldots, \boldsymbol{x}_{T,N})^T.$$

The notation A_H means that the matrix A has columns restricted to the set H, while $\boldsymbol{x}^T_{t,j}$ is the jth row of $\boldsymbol{X}_t$.

Assumption M1 ensures that our model has a reduced form

$$y_t = \boldsymbol{\Pi}\boldsymbol{\mu} + \boldsymbol{\Pi}\boldsymbol{W}_1 y_{t-1} + \cdots + \boldsymbol{\Pi}\boldsymbol{W}_p y_{t-p} + \boldsymbol{\Pi}\boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\Pi}\boldsymbol{\epsilon}_t, \quad \boldsymbol{\Pi} = (\boldsymbol{I}_N - \boldsymbol{W})^{-1},\ t = 1, \ldots, T.$$

The matrix $\boldsymbol{\Pi}$ exists with the assumption $\sum_{i=1}^M |\delta_{0i}| < 1$. Since row-standardization means $\left\|\boldsymbol{W}_{0i}\right\|_\infty = 1$, the condition $\sum_{j=1}^p \sum_{i=1}^M |\delta_{ji}| + \sum_{s\in S} |\beta_s| < 1$ implies that each $\left\|\boldsymbol{W}_j\right\|_\infty < \sum_{i=1}^M |\delta_{ji}| \left\|\boldsymbol{W}_{0i}\right\|_\infty < 1$. At the same time, without loss of generality assuming $S = \phi$ and writing the model as

$$\boldsymbol{\Phi}(L) y_t = \boldsymbol{\Pi}\boldsymbol{\mu} + \boldsymbol{\Pi}\boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\Pi}\boldsymbol{\epsilon}_t, \quad \boldsymbol{\Phi}(L) = (\boldsymbol{I}_N - \boldsymbol{\Pi}\boldsymbol{W}_1 L - \cdots - \boldsymbol{\Pi}\boldsymbol{W}_p L^p),$$

where L is the lag operator, then stationarity is ensured if $\det(\boldsymbol{\Phi}(z)) = 0$ has all roots lying outside the unit circle. This is ensured by the condition $\sum_{j=1}^p \sum_{i=1}^M |\delta_{ji}| + \sum_{s\in S} |\beta_s| < 1$, which is thus a sufficient condition for stationarity. In practice, we implement these restrictions when finding $\widetilde{\boldsymbol{\delta}}$ in Section 1.2.3.

The Assumption M2 and M3 are also used in Lam and Souza (2018). The independence between $\{\boldsymbol{B}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ in M2 ensures that $\{\boldsymbol{B}_t\}$ serves a function similar to an instrument for model (1.2). More details about the dependence between $\{\boldsymbol{B}_t\}$ and $\{\boldsymbol{X}_t\}$ are shown in Assumption R3 and R4 in Section 1.8.1.

The tail condition in M2 implies that all the random variables involved are with sub-exponential tails, which is a relaxation to strict normality. The different constants for $\{\boldsymbol{B}_t\}$, $\{\boldsymbol{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ process are finally dominated by the maximum one among them. Therefore, to simplify, we assume they are same in our proof.

Similar to Definition 3 in Wu (2005) for stability measurement, $\Theta^x_{m,2w} \leq Cm^{-\alpha}$ in M3 essentially means that the strongest serial dependence for the x_{tj}'s with at least m time units apart is decaying polynomially as m increases. It allows for the application of a Nagaev-type inequality in Lemma 1 in the Section 1.8 for our results to hold.

1.3.2 Identification of the model

To explain condition M4 about identification of the model more specifically, we assume that we have two sets of parameters $(\boldsymbol{\beta}^*, \boldsymbol{\delta}^*)$ and $(\boldsymbol{\beta}^o, \boldsymbol{\delta}^o)$ that satisfy model (1.2). Then we have

$$\mathbf{0} = \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}_H (\boldsymbol{\delta}^*_H - \boldsymbol{\delta}^o_H) + \boldsymbol{B}^T \boldsymbol{X}_{\boldsymbol{\beta}^* - \boldsymbol{\beta}^o} \mathrm{vec}(\boldsymbol{I}_N),$$

and we can write

$$\begin{aligned} T^{-1/2} \boldsymbol{B}^T \boldsymbol{X}_{\boldsymbol{\beta}^* - \boldsymbol{\beta}^o} \mathrm{vec}(\boldsymbol{I}_N) &= N^{-a/2} \begin{pmatrix} T^{-1} \sum_{t=1}^T (\boldsymbol{B}_t - \bar{\boldsymbol{B}}) \boldsymbol{\zeta} \mathbf{x}_{t,1}^T (\boldsymbol{\beta}^* - \boldsymbol{\beta}^o) \\ \vdots \\ T^{-1} \sum_{t=1}^T (\boldsymbol{B}_t - \bar{\boldsymbol{B}}) \boldsymbol{\zeta} \mathbf{x}_{t,N}^T (\boldsymbol{\beta}^* - \boldsymbol{\beta}^o) \end{pmatrix} \\ &= T^{-1/2} \boldsymbol{B}^T \tilde{\boldsymbol{X}} (\boldsymbol{\beta}^* - \boldsymbol{\beta}^o), \end{aligned}$$

so that

$$[T^{-1/2} \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}_H \;\; T^{-1/2} \boldsymbol{B}^T \tilde{\boldsymbol{X}}] \begin{pmatrix} \boldsymbol{\delta}^*_H - \boldsymbol{\delta}^o_H \\ \boldsymbol{\beta}^* - \boldsymbol{\beta}^o \end{pmatrix} = \mathbf{0}.$$

Hence taking expectation and multiplying $\boldsymbol{O}^T$ on both sides and then $(\boldsymbol{O}^T \boldsymbol{O})^{-1}$, with condition M4, we can show that $\boldsymbol{\delta}^*_H = \boldsymbol{\delta}^o_H$ and $\boldsymbol{\beta}^* = \boldsymbol{\beta}^o$. Since $|H| \leq M(p+1)$

and K are finite and N^2 is much larger than $|H| + K$, assuming $\boldsymbol{O}$ whose size is $N^2 \times (|H| + K)$ has full rank is reasonable.

1.3.3 Main results

To show the main theorem, we define $\lambda_T = cT^{-1/2}\log^{1/2}(T \vee N)$, where $c > 0$ is a constant. In all theorems presented here, we assume that $\alpha \geq 1/2 - 1/w$ in Assumption M3, which is part of the further assumptions listed in the Section 1.8.

Theorem 1. *Let the assumptions in Section 1.3.1 and in Theorem 5 hold. The estimators $\hat{\boldsymbol{\delta}}$ in (1.5) and $\hat{\boldsymbol{\beta}}$ in (1.6) satisfy*

$$\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1 = O_P(\lambda_T N^{-1/2+1/2w}) \quad \textit{and}$$
$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 = O_P(\lambda_T N^{-1/2+1/2w}).$$

Same as Lam and Souza (2018), $w > 2$ is assumed in M3, the above immediately implies $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1, \|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1 \to 0$ in probability. First, it makes sense as $T \to \infty$, as T is the sample size in our setting. It also makes perfect sense as $N \to \infty$ since we are accumulating more information cross-sectionally for the finite-sized parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ as N goes to infinity. Second, $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\beta}}$ converge to the corresponding true values at the same rate which matches the fact shown in (1.6) that $\widehat{\boldsymbol{\beta}}$ is expressed by $\widehat{\boldsymbol{\delta}}$. Then we present the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$ in the following two theorems.

Theorem 2. *Let the assumptions in Section 1.3.1 and in Theorem 5 hold. Moreover, use the predictive dependence measures defined in Wu (2005) as*

$$\boldsymbol{P}_0^b(B_{t,qk}) = \mathbb{E}(B_{t,qk}|\mathcal{G}_0) - \mathbb{E}(B_{t,qk}|\mathcal{G}_{-1}), \quad \boldsymbol{P}_0^\epsilon(\epsilon_{t,qk}) = \mathbb{E}(\epsilon_{t,qk}|\mathcal{H}_0) - \mathbb{E}(\epsilon_{t,qk}|\mathcal{H}_{-1}),$$

where $\mathcal{G}_t$ and $\mathcal{H}_t$ are defined in Section 1.3. Assume

$$\sum_{t\geq 0} \max_{1\leq q\leq N} \max_{1\leq k\leq K} \|\boldsymbol{P}_0^b(B_{t,qk})\| \leq \infty, \quad \sum_{t\geq 0} \max_{1\leq j\leq N} \|\boldsymbol{P}_0^\epsilon(\epsilon_{tj})\| \leq \infty.$$

Then we have

$$T^{1/2}\boldsymbol{\Sigma}_1^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_K),$$

where
$\boldsymbol{\Sigma}_1 = \boldsymbol{M}_1 \sum_{\tau\in\mathbb{Z}} \mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_{t+\tau}^T \boldsymbol{B}_{t+\tau}) \boldsymbol{M}_1^T$ *and* $\boldsymbol{M}_1 = (\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t))^{-1}\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)$.

Theorem 3. *Let the assumptions in Section 1.3.1 and in Theorem 5 hold. Assume that the predictive dependence measures* $\boldsymbol{P}_0^b(B_{t,qk})$ *and* $\boldsymbol{P}_0^\epsilon(\epsilon_{t,qk})$ *are as defined in Theorem 2 with the same assumptions applied. Then*

$$T^{1/2}\boldsymbol{\Sigma}_2^{-1/2}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}) \xrightarrow{\mathcal{D}} \boldsymbol{N}(\mathbf{0}, \boldsymbol{I}_{M(p+1)}),$$

where $\boldsymbol{\Sigma}_2 = \boldsymbol{M}_2(\mathbf{S}_1+\mathbf{S}_2-\mathbf{S}_3-\mathbf{S}_3^T)\boldsymbol{M}_2^T$, *and*

$$\begin{aligned}
\mathbf{S}_1 &= \sum_{\tau\in\mathbb{Z}} \mathbb{E}(\boldsymbol{M}\boldsymbol{B}_{t+\tau}^T\boldsymbol{\epsilon}_{t+\tau}\boldsymbol{\epsilon}_t^T\boldsymbol{B}_t^T\boldsymbol{M}^T),\\
\mathbf{S}_2 &= \sum_{\tau\in\mathbb{Z}} \left[\mathbb{E}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T)\otimes\mathbb{E}(\boldsymbol{B}_t\boldsymbol{\zeta}\boldsymbol{\zeta}^T\boldsymbol{B}_{t+\tau}^T)^T\right],\\
\mathbf{S}_3 &= \sum_{\tau\in\mathbb{Z}} \mathbb{E}(\boldsymbol{M}\boldsymbol{B}_{t+\tau}^T\boldsymbol{\epsilon}_{t+\tau}(\mathrm{vec}(\boldsymbol{B}_t\boldsymbol{\zeta}\boldsymbol{\epsilon}_t^T))^T),\\
\boldsymbol{M}_2 &= \{(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})^T(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})\}^{-1}(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})^T,\ \textit{with}\\
\boldsymbol{H}_{10} &= [\boldsymbol{I}_N\otimes\mathbb{E}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\zeta}y_t^T),\ldots,\boldsymbol{I}_N\otimes\mathbb{E}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\zeta}y_{t-p}^T)]\boldsymbol{V},\\
\boldsymbol{H}_{20} &= \boldsymbol{M}[\mathbb{E}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T\boldsymbol{W}_{01}y_t),\ldots,\mathbb{E}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T\boldsymbol{W}_{0M}y_t),\ldots,\\
&\qquad \mathbb{E}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T\boldsymbol{W}_{01}y_{t-p}),\ldots,\mathbb{E}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T\boldsymbol{W}_{0M}y_{t-p})],
\end{aligned}$$

where $\boldsymbol{M} = \mathbb{E}(\boldsymbol{X}_t\otimes(\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\zeta})\left[\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t)\right]^{-1}\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)$.

Similar to the convergence property found in Theorem 1, these two theorems are also important, since they provide the tools for practical data analysis such as hypothesis testing and confidence intervals construction.

In practice, we need to consider the infinite summation over τ and the high dimension matrix calculation when the above asymptotic normality results are applied. More specifically, when calculating $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, we calculate sample means to replace the corresponding expectations. For the infinite summations in τ in $\mathbf{S}_1$ to $\mathbf{S}_3$, we check if the matrix at a particular τ has very small elements overall by calculating Frobenius norm. If so, we discard the whole matrix and all the matrices beyond this particular τ. In the real data analysis in Section 1.6, we find that we always discard those with $\tau \geq 5$. See Section 1.4.1 for further treatments regarding the estimation of the high dimensional matrices $\mathbf{S}_1$ to $\mathbf{S}_3$.

Theorem 4. *(Oracle property for* $\tilde{\boldsymbol{\delta}}$*) Let the assumptions in Section 1.3.1 and in*

Theorem 5 hold. Then as $T, N \to \infty$, with probability approaching 1,

$$\text{sign}(\tilde{\boldsymbol{\delta}}_H) = \text{sign}(\boldsymbol{\delta}_H), \quad \tilde{\boldsymbol{\delta}}_{H_c} = 0,$$

where $H = \{\ell : \delta_\ell \neq 0\}$ and $\ell = 1, \ldots, M(p+1)$. Moreover, let the predictive dependence measures $\boldsymbol{P}_0^b(B_{t,qk})$ and $\boldsymbol{P}_0^\epsilon(\epsilon_{t,qk})$ be as defined in Theorem 2 with the same assumptions applied. Then

$$T^{1/2}\boldsymbol{\Sigma}_3^{-1/2}(\tilde{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H) \xrightarrow{\mathcal{D}} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_{|H|}),$$

where $\boldsymbol{\Sigma}_3 = \boldsymbol{M}_3(\mathbf{S}_1 + \mathbf{S}_2 - \mathbf{S}_3 - \mathbf{S}_3^T)\boldsymbol{M}_3^T$, and $\boldsymbol{M}_3 = \{(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H\}^{-1}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T$

As the usage of adaptive LASSO, we can build the oracle property for $\widetilde{\boldsymbol{\delta}}$ to carry out the selection of the specified spatial weight matrices by the penalized estimator $\widetilde{\boldsymbol{\delta}}$, and the usual inferences on the non-zero elements in $\widetilde{\boldsymbol{\delta}}$. The practical performances of these estimators and the asymptotic normality results are presented in Section 1.5.

1.4 Practical Implementation

1.4.1 Regularized matrix estimation of $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$

As discussed in Theorem 3 and 4, the definitions of $\mathbf{S}_1$ to $\mathbf{S}_3$ involve some high dimensional matrices to be estimated. Since $\mathbf{S}_1$ to $\mathbf{S}_3$ are in fact all $N^2 \times N^2$, in this paper we regularize $\mathbf{S}_1$ and $\mathbf{S}_3$ by banding them directly (see Bickel and Levina (2008) for more details). As for the banding width used in our simulation and real data analysis, we apply the 5-fold cross-validation procedure suggested in Bickel and Levina (2008). It is found that that retaining only two off-diagonals (two upper and two lower, while setting 0 in all other off-diagonals) when $\tau = 0$, and retaining only one when $|\tau| \geq 1$ in the infinite summations in $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$ achieves good results when N is moderate to large. Again similar to the discussion after Theorem 3, when $|\tau| \geq 5$, we set the matrices inside the summations in the definitions of $\mathbf{S}_1$ to $\mathbf{S}_3$ to exactly zero. For $\mathbf{S}_2$, there are two $N \times N$ matrices $\mathbb{E}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T)$ and $\mathbb{E}(\boldsymbol{B}_t\boldsymbol{\zeta}\boldsymbol{\zeta}^T\boldsymbol{B}_{t+\tau}^T)$. We band them separately, again retaining only two off-diagonals each when $\tau = 0$, and only one when $|\tau| \geq 1$. It is also suggested for users to apply this idea in practice, as our method performs very well on both the simulated and real data as shown in Section (1.5) and (1.6).

1.4.2 Choice of the number of time lags p, and γ_T

In our analysis, we assume that p in model (1.1) is prescribed and finite. Same as Lam and Souza (2018), for practical data analysis, we choose p by minimizing the following BIC criterion:

$$\text{BIC}(p) = \log\big(N^{-1}\|\boldsymbol{B}^T y - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}\hat{\boldsymbol{\delta}} - \boldsymbol{B}^T\boldsymbol{X}_{\hat{\boldsymbol{\beta}}}\text{vec}(\boldsymbol{I}_N)\|^2\big) + p\frac{\log T}{T}\log(\log T), \quad (1.8)$$

which follows the one in Wang et al. (2009). Wang et al. (2009) shows that this modified BIC is consistent in model selection, which is correct regardless of whether the dimension of the true model is finite or diverging. Section 1.5 also shows the desirable performance in practice. Note that in the definition of $\boldsymbol{B}$, there is a rate a which is unknown. However, because of the logarithmic operation in the first term in BIC(p), the value of a does not change where the minimum of BIC(p) is achieved.

For the choice of γ_T, we use the BIC criterion above, but with $\widehat{\boldsymbol{\delta}}$ replaced by $\tilde{\boldsymbol{\delta}}$, so that we are effectively choosing p and γ_T together. Cross validation can also work for the choice of the number of time lags p and γ_T, but we apply the above BIC criterion for the computational efficiency.

1.4.3 Choice of $\boldsymbol{\zeta}$ in $\boldsymbol{B}$

We have set $\boldsymbol{\zeta} = K^{-1}1_K$ as fixed in the definition of $\boldsymbol{B}$ in Section 1.2.2. In fact this can be estimated to provide maximal correlation between $\boldsymbol{B}$ and the response variable y_t through two-stage least squares. Consider the model

$$y_t = \boldsymbol{\alpha} + \boldsymbol{B}_t\boldsymbol{\zeta} + \boldsymbol{v}_t,$$

where $\boldsymbol{\alpha}$ is an $N \times 1$ vector of unknown coefficients, and $\boldsymbol{\zeta}$ is the $K \times 1$ vector of coefficients we want to estimate. To get $\hat{\boldsymbol{\zeta}}$, we can consider the problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\zeta}} \sum_{t=1}^{T} \|y_t - \boldsymbol{\alpha} - \boldsymbol{B}_t\boldsymbol{\zeta}\|^2,$$

with solution

$$\hat{\boldsymbol{\zeta}} = \Big(\sum_{t=1}^{T}(\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T(\boldsymbol{B}_t - \bar{\boldsymbol{B}})\Big)^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T(\mathrm{y}_t - \bar{y}).$$

Implementing this does not change our proofs, since it is easy to show that $\left\|\hat{\boldsymbol{\zeta}}\right\|_1 = O_P(1)$, which substitutes $\left\|\widehat{\boldsymbol{\zeta}}\right\|_1 = 1$ in all of our proofs. We have tried this in our simulations and real data analysis, and the practical differences between using this and $\boldsymbol{\zeta} = K^{-1}1_K$ is negligible.

1.5 Simulation Experiments

1.5.1 Setting and results

To generate y_t through model (1.1), we generate $\boldsymbol{X}_t$ by using $\text{vec}(\boldsymbol{X}_t) = 0.2 \cdot 1_K \otimes \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_t^{\boldsymbol{X}}$ with $K = 3$, where $\boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{I}_N)$ is the innovation series for model (1.1), with the $\boldsymbol{\epsilon}_t$'s being independent of each other. The $\boldsymbol{\epsilon}_t^{\boldsymbol{X}}$'s are independent of each other and of other variables, with $\boldsymbol{\epsilon}_t^{\boldsymbol{X}} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$, and

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \begin{bmatrix} 2\boldsymbol{I}_N & 0.5\boldsymbol{I}_N & 0.5\boldsymbol{I}_N \\ 0.5\boldsymbol{I}_N & 2\boldsymbol{I}_N & 0.5\boldsymbol{I}_N \\ 0.5\boldsymbol{I}_N & 0.5\boldsymbol{I}_N & 2\boldsymbol{I}_N \end{bmatrix}.$$

Since $\boldsymbol{X}_t$ depends on $\boldsymbol{\epsilon}_t$, we set $\boldsymbol{B}_t$ to be such that $\text{vec}(\boldsymbol{B}_t) = 0.7\boldsymbol{\epsilon}_t^{\boldsymbol{X}} + \boldsymbol{\epsilon}_t^{\boldsymbol{B}}$, where the $\boldsymbol{\epsilon}_t^{\boldsymbol{B}}$'s are drawn independently from the same distribution as $\boldsymbol{\epsilon}_T^{\boldsymbol{X}}$, and they are independent of all other variables.

We set $M = 3$ and $p = 2$ for the model. Each element of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ is generated independently from the uniform distribution $\boldsymbol{U}(0, 1)$. Elements in $\boldsymbol{\delta}$ are then randomly chosen to be 0 while maintaining $p = 2$. To make sure the stationarity of $\{y_t\}$, every element in $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ is then divided by 1.1 times the absolute sum of all elements in $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ respectively.

For the $M = 3$ specified spatial weight matrices, to facilitate stationarity of the model, we construct each $\boldsymbol{W}_i$ such that only the first three off-diagonals (upper and lower) have non-zero elements. This way, as N increases, we can still control the eigenvalues of $\boldsymbol{W}_i$ to be less than 1 in magnitude. In another setting, we generate an orthogonal matrix $\boldsymbol{V}_i$ and a diagonal matrix $\mathbf{D}_i$ with all values in $\mathbf{D}_i$ to be less than 1 in magnitude, such that $\boldsymbol{W}_i = \boldsymbol{V}_i\mathbf{D}_i\boldsymbol{V}_i^T$. Ultimately, both settings achieves very similar results, and hence we only show the results of the former setting.

We repeat our simulations for 500 times, and report the averaged L_1-error for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$ (i.e., respectively, $\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_1/3 = \sum_{i=1}^3 |\widehat{\beta}_i - \beta_i|/3$ and $\left\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right\|_1/9$) in Figure

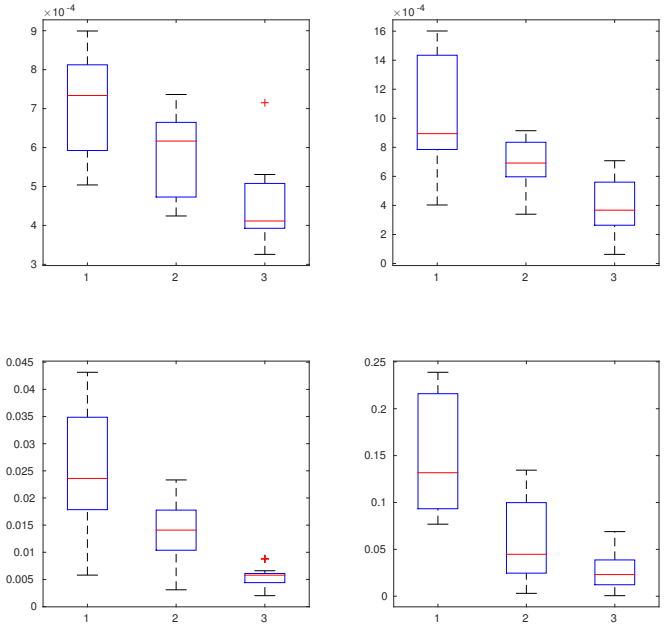


Figure 1.1: Boxplots of averaged L_1 errors. Upper row: $\sum_{i=1}^{3} |\widehat{\beta}_i - \beta_i|/3$. Bottom row: $\left\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\right\|_1/9$. Left column (from left to right): $N = 40, 80, 120, T = 60$. Right column (from left to right): $T = 40, 80, 120, N = 60$.

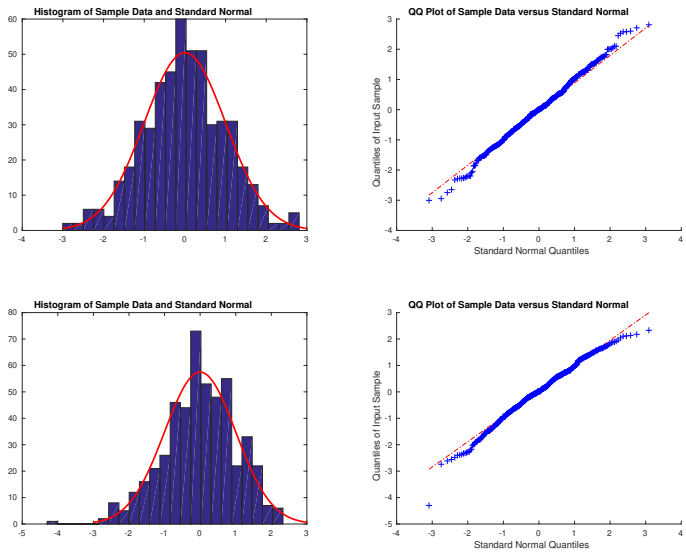


Figure 1.2: Histograms and normal probability plots for standardized $\hat{\beta}_1$ (upper row) and $\hat{\delta}_{1,3}$ (lower row) with $N = T = 80$. Standardization used respectively the asymptotic results from Theorem 2 and 3.

1.1, which illustrates the convergence of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$ respectively as N, T or both gets larger.

Next we consider the asymptotic normality of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$. We choose $\widehat{\beta}_1$ and $\widehat{\delta}_{1,3}$ with $(N, T) = (80, 80)$ as examples for illustration. For each simulation, we construct $\widehat{\beta}_1$ and $\widehat{\delta}_{1,3}$, and standardize them according to the asymptotic results in Theorem 2 and Theorem 3 respectively. Figure 1.2 shows histograms and normal probability plots of the standardized estimators. They both show good fit for a standard normal distribution. It means that the asymptotic variance formulae in Theorem 2 and 3 are reliable for inference, and the way that we estimate any high dimensional covariance matrices mentioned in Section 1.4.1 helps in achieving an accurate estimation of the covariance matrices for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$. We actually get very similar good fits for the non-zero components of $\tilde{\boldsymbol{\delta}}$, showing the asymptotic normality in Theorem 4 is reliable as well. The results are omitted here to save space.

On top of asymptotic normality, $\tilde{\boldsymbol{\delta}}$ also enjoys sign consistency as shown in Theorem 4. We illustrate the selection consistency of $\tilde{\boldsymbol{\delta}}$ in practice by calculating the specificity (i.e., proportion of correctly identified zeros) and the sensitivity (i.e., proportion of correctly identified non-zeros) of $\widetilde{\boldsymbol{\delta}}$. Table 1.1 shows that at various combinations of (N, T), the sensitivity and specificity are all 100%, showing perfect identifications of zeros and non-zeros. The table also shows the decreasing error for $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\delta}}$ as N or T increases.

		$T = 40$	$T = 80$	$T = 120$
$N = 60$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$	9.06(3.58)	6.26(1.04)	3.16(0.58)
	$\|\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1$	0.13(0.05)	0.05(0.04)	0.02(0.02)
	$\widetilde{\boldsymbol{\delta}}$ Specificity	100%(0)	100%(0)	100%(0)
	$\widetilde{\boldsymbol{\delta}}$ Sensitivity	100%(0)	100%(0)	100%(0)
		$N = 40$	$N = 80$	$N = 120$
$T = 60$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$	7.60(0.89)	6.24(0.77)	3.51(0.65)
	$\|\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1$	0.02(0.01)	0.01(0.00)	0.00(0.00)
	$\widetilde{\boldsymbol{\delta}}$ Specificity	100%(0)	100%(0)	100%(0)
	$\widetilde{\boldsymbol{\delta}}$ Sensitivity	100%(0)	100%(0)	100%(0)

Table 1.1: Mean L_1 error for $\widehat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\delta}}$. Standard deviations are shown in brackets. Sensitivity and specificity of $\widetilde{\boldsymbol{\delta}}$ are also shown for various combinations of T, N. The values of $\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_1$ are multiplied by 10^4.

1.5.2 Cross-sectional dependence in the innovation

Same as the simulation setting used in Section 1.5.1, a strong cross-sectional dependent error is also considered in our simulation, where its i,jth element covariance

matrix $\mathbf{\Sigma}^{\epsilon}_{i,j} = \alpha^{|i-j|}$ and $\alpha = 0.8$. The results about the proposed estimators are shown in Table 1.2 when different combinations of sample size T and panel dimension are used. First, the spatial weight matrix selection results are still perfect, as all values of $\widetilde{\boldsymbol{\delta}}$ Specificity and Sensitivity are 100% for different T and N. Secondly, the mean L_1 errors for $\widehat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\delta}}$ become significantly large than the simulation where there is no cross=sectional dependence in the innovation process generation. However, with the increase of sample size T or panel dimension N, the mean L_1 errors for $\widehat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\delta}}$ decrease.

		$T = 40$	$T = 80$	$T = 120$
$N = 80$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$	45.3(11.9)	34.6(12.4)	26.66(10.5)
	$\|\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1$	2.02(0.70)	1.67(0.42)	1.51(0.54)
	$\widetilde{\boldsymbol{\delta}}$ Specificity	100%(0)	100%(0)	100%(0)
	$\widetilde{\boldsymbol{\delta}}$ Sensitivity	100%(0)	100%(0)	100%(0)
		$N = 40$	$N = 80$	$N = 120$
$T = 80$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$	47.7(22.7)	34.6(12.4)	28.0(10.8)
	$\|\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1$	1.88(0.61)	1.67(0.42)	1.63(0.45)
	$\widetilde{\boldsymbol{\delta}}$ Specificity	100%(0)	100%(0)	100%(0)
	$\widetilde{\boldsymbol{\delta}}$ Sensitivity	100%(0)	100%(0)	100%(0)

Table 1.2: Mean L_1 error for $\widehat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\delta}}$ when innovation process contains cross-sectional dependence. Standard deviations are shown in brackets. Sensitivity and specificity of $\widetilde{\boldsymbol{\delta}}$ are also shown for various combinations of T, N. The values of $\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_1$ are multiplied by 10^4.

1.5.3 Performance of BIC for choosing p

To examiner the performance of the BIC defined in (1.8), we run our simulations 100 times for each particular (N, T) combination using the same setting as before, except that each time p is randomly generated from 1 to 7. With each simulation, we construct the positive selection rate (PSR) and the false discovery rate (FDR), defined as

$$\text{PSR} = \frac{\sum_{j=1}^{100} |s_j^* \cap s_{0,j}|}{\sum_{j=1}^{100} |s_{0,j}|}, \quad \text{FDR} = \frac{\sum_{j=1}^{100} |s_j^* \cap s_{0,j}^c|}{\sum_{j=1}^{100} |s_j^*|},$$

where $s_{0,j}$ represents the index set for all δ_{ir} that should be included in the model at the jth repetition. Since we do not set δ_{ir} to be exactly 0 in this experiment, we have $|s_{0,j}| = pM = 3p$, where p is in fact different for different j. The set s_j^* is the index set for all $\widehat{\delta}_{ir}$ estimated when p is estimated as p^*. Clearly, if $p^* \leq p$, then $|s_j^* \cap s_{0,j}| = |s_j^*|$ and $|s_j^* \cap s_{0,j}^c| = 0$, meaning we may not be having the whole true set $s_{0,j}$ but we do not falsely "discover" something that is not in $s_{0,j}$. On the

other hand, if $p^* > p$, then $|s_j^* \cap s_{0,j}| = |s_{0,j}|$ and $|s_j^* \cap s_{0,j}^c| > 0$, meaning we have included all that are in $s_{0,j}$, but we have falsely "discovered" something outside of $s_{0,j}$. Hence in a sense, PSR measures an average number of times where we do not underestimate p, while FDR measures an average number of times we overestimate p. Ideally, we want PSR=100% while FDR = 0%. These two measures are also used in Chen and Chen (2008) and Chen and Chen (2012) in different contexts.

		$T = 40$	$T = 50$	$T = 60$
$N = 50$	PSR	100.00%	100.00%	98.00%
	FDR	2.00%	0.00%	0.00%
		$N = 40$	$N = 50$	$N = 60$
$T = 50$	PSR	98.00%	100.00%	100.00%
	FDR	0.00%	0.00%	2.00%

Table 1.3: Positive selection rate (PSR) and false discovery rate (FDR) for the choice of p using BIC defined in (1.8).

Table 1.3 shows the results. Our BIC definitely performs very well with PSR almost always equal 100% and FDR 0% in various (N, T) combinations.

1.6 Analysis of Stock Return Data

Spatial lag model has been widely applied to economic or geographic data, yet financial data is rarely analyzed using spatial econometrics tools. We illustrate the performance of our model using the daily log-returns of 32 important stocks shown in the following table in the Euro Stoxx 50 and S&P 500 in 2015. Our aim is to analyze the spatial interactions of these stocks and to see how different macroeconomic and financial indicators affect the dynamics of the returns.

Arnold et al. (2013a) illustrates, with the help of a spatial lag model, that the stocks belonging to the same country or the same industry are more related to each other, in the sense that spatial interactions of the log-returns are stronger. They analyze the Euro Stoxx 50 stock returns using a combination of three adjacency matrices as an estimator for the spatial weight matrix in their model. The first one being the weight of the stocks in Euro Stoxx 50, and the second and third ones being the adjacency matrices corresponding to the same industry and to the same country, respectively. They found that all these matrices contribute to the final spatial weight matrix in their model, and improves risk estimation in a portfolio allocation exercise. However, no selection and inferences on the estimated parameters are performed due to the lack of regularization and asymptotic results for the estimators.

France	Alstom, Total, BNP, Scociete, Sanofi, Carrefour, LVMH, Vivendi
Germany	Daimler, Allianz, Deutsche Bank
Italy	ENEL, ENI, Intesa, Unicredit, Tele Italy
Spain	Repsol, Banco, Telefonica
US	GM, PG, Nextera, American Express, Citi, Wells Frago, Amgen, Gilead, Johnson, Costco, Home, Centurylink, Verizon
Energy	Alstom, Total, ENEL, ENI, Repsol, PG, Nextera
Finance	BNP, Scociete, Allianz, Deutsche Bank, Intesa, Unicredit, Banco, American Express, Citi, Wells Fargo
Pharmacy	Sanofi, Amgen, Gilead, Johnson
Retails	Carrefour, LVMH, Costco, Home
Telecom	Vivendi, Tele Italy, Telefonica, Centurylink, Verizon
Auto	Daimler, GM

To fill in this gap and generalize on their model, we include four types of spatial weight matrix specifications instead of only three matrices as in Arnold et al. (2013a). The first type is on the physical distance d_{ij} between city i and j where the headquarters of the stocks' associated companies are built. As stock market is significantly affected by the local economy and, in spatial economy research, physical distance is commonly used. In our case, three specified spatial weight matrices with elements $1/d_{ij}, 1/d_{ij}^2$ and $1/d_{ij}^3$ are included for selection. The second to fourth types coincide with the three matrices specified in Arnold et al. (2013a). Namely, one contains the weight of stocks in Euro Stoxx 50 or S&P 500, and the remaining two having (i, j)th element equal to 1 if the corresponding stocks belong to the same industry or country respectively. This way, we have $M = 6$ specified spatial weight matrices for selection in our model. We have done row standardization on all of these six matrices.

As for the covariates $\boldsymbol{X}_t$, we use the Fama-French three factors (excess return = market return - risk free rate, SMB = Small (market capitalization) Minus Big, HML = High (book-to-market ratio) Minus Low), national stock index (S&P 500, CAC40, DAX, IBEX or MIB) and the corresponding European or US industry index for each stock. Hence $K = 5$, and we are treating these as exogenous covariates, so we set $\boldsymbol{B}_t = \boldsymbol{X}_t$, the same as the covariates. Minimizing the BIC defined in (1.8) results in $p = 1$.

Table 1.4 shows the values of $\tilde{\boldsymbol{\delta}}$. Clearly, stock weight in their respective market indices do not contribute to the two spatial weight matrices $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$. However, the adjacency matrices for country and industry do contribute to both of the spatial weight matrices. For physical distance, clearly, a traditional approach where one

	$1/d$	$1/d^2$	$1/d^3$	Stock weight	Country	Industry
$\widetilde{\boldsymbol{\delta}}_{0i}$	-0.0052	0.0811	-0.3880	0	0.0001	0.3122
	(0.0015)	(0.0036)	(0.0497)	(—)	(10^{-5})	(0.0346)
$\widetilde{\boldsymbol{\delta}}_{1i}$	0	0	-0.0612	0	2.22×10^{-5}	0.0610
	(—)	(—)	(0.0062)	(—)	(6.1×10^{-6})	(0.0051)

	Market excess return	SMB	HML
$\hat{\boldsymbol{\beta}}$	1.516(0.616)	−4.884(2.662)	1.869(0.841)
	National Index	Industry Index	
$\hat{\boldsymbol{\beta}}$	14.788(5.563)	19.746(10.274)	

Table 1.4: The values of $\widetilde{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\beta}}$, where $p = 1$ and $\gamma_T = 1.6438$ are chosen by minimizing the BIC defined in (1.8). Estimated standard deviations are in brackets. All values associate with $\widehat{\boldsymbol{\beta}}$ are multiplied by 10^6.

chooses a distance $1/d$, $1/d^2$ or $1/d^3$ for the spatial weight matrix would fail, since it is clear that all three specified spatial weight matrices are significant and cannot be omitted for $\boldsymbol{W}_0$. Only the one for $1/d^3$ is significant to $\boldsymbol{W}_1$ though. In the same table, we can see that all factors in $\boldsymbol{X}_t$ are at least marginally significant, with national and industry indices play a more important role practically than the Fama-French three factors.

Figure 1.3 shows the heat map of the spatial weight matrices $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$. It is clear that there are some block patterns in these matrices, which mainly represent stocks in the same country or industry. Meanwhile, they are related strongly with each other in general if they are all from Europe or US, with France and Italy showing strong connections. It is interesting to note that the ninth stock Daimler, and the twentieth stock GM, are related to each other (two bright yellow dots on both $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$), although they belong to Germany and US auto-industry respectively. Since Daimler owns part of GM by spin-offs, the relation itself is not surprising. However, it means that our method of taking linear combination of different specified spatial weight matrices can indeed reflect a general pattern of spatial interactions. In $\boldsymbol{W}_1$, we can also find some blocks for stocks in Germany and Spain.

1.7 Conclusion

In this chapter, we derive the properties of profiled least square estimation of the time-space dynamic model that can contain contemporaneous spatial effect, individual time-lagged effects and spatial time-lagged effects. The convergence properties

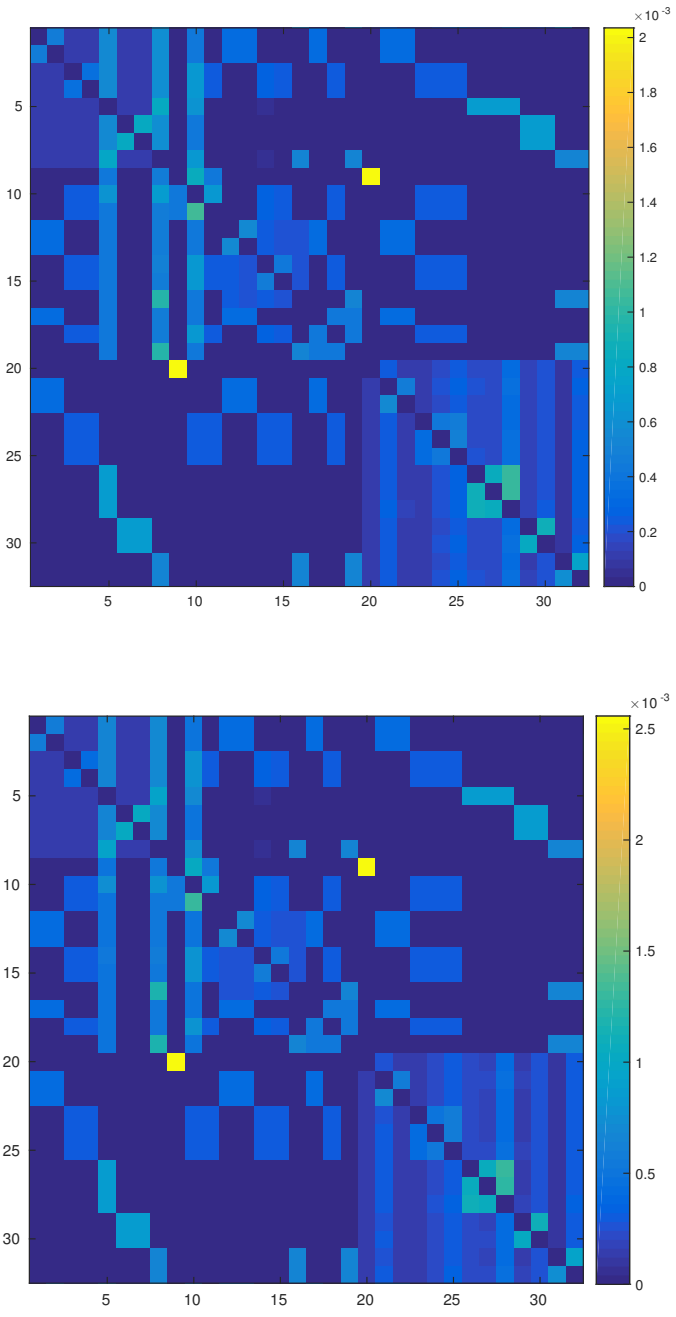


Figure 1.3: Upper: The estimate of $\boldsymbol{W}_0$. Lower: The estimate of $\boldsymbol{W}_1$. From 1 to 32, the stocks are Alstom, Total, BNP, Scociete, Sanofi, Carrefour, LVMH, Vivendi, Daimler, Allianz, Deutsche Bank, ENEL, ENI, Intesa, Unicredit, Tele Italy,Repsol, Banco, Telefonica, GM, PG, Nextera, American Express, Citi, Wells Frago, Amgen, Gilead, Johnson, Costco, Home, Centurylink and Verizon respectively.

and the asymptotic normality results are built in the case when both the sample size T and panel dimension N go to infinity. Using instrument-like variables, the inconsistency from innate endogeneity in least square estimators can be fixed.

Our model uses different linear combination of a set of specified spatial weight matrices for different spatial effects to avoid the misspecification. However, it is highly likely that some irrelevant spatial weight matrices are considered into the model, which may cause a new bias for the spatial weight matrix estimation. A further selection on spatial weight matrices included into our model is applied by adaptive LASSO proposed in Zou (2006), which can reflect which spatial weight matrices truly contribute which ones do not.

As for the prediction ability of our model, it is easy to have the predictive value by:

$$\hat{y}_t = (\boldsymbol{I} - \hat{\boldsymbol{W}}_0)^{-1}(\hat{\boldsymbol{\mu}} + \hat{\boldsymbol{W}}_1 y_{t-1} + \cdots + \hat{\boldsymbol{W}}_p y_{t-p} + \boldsymbol{X}_t\hat{\boldsymbol{\beta}}).$$

Same as the most of VAR (Vector AutoRegression) model, our model dose not have a good predictive ability when the panel dimension N is large. The estimated inverse matrix in the above predictive model do not perform well, which causes the predictive values inaccurate. However, the main goal of this Chapter is to construct the spatial weight matrix by high order spatial autoregression model and do a selection on the candidates of specified spatial weight matrices. We can leave the forecasting problem as a future work.

For the future study, first of all, the fixed M assumed in the proposed method can be extended to infinite M to reflect the practical reality that richness of a parametric model often deepens with sample size. Furthermore, it may be of interest to increase the efficiency of spatial weight matrix estimation. As known, adaptive LASSO is good in spatial weight matrices selection, however, it introduces some bias into the spatial weight matrix estimation as the sacrifice of sparseness. In Lam and Souza (2015c), adding a potentially sparse adjustment matrix for contemporaneous spatial effect is discussed, which can be extended to dynamic spatial model we proposed in this chapter. In the case of p is not large, quasi-maximum likelihood estimation can also be applied, especially when the instrument-like variable is not available. The last but not the least, we can still apply the proposed method but replace the L_1 penalty by ridge regularization, which performs better in estimation but not in selection. We will investigate these approaches in a future work.

1.8 Proof

1.8.1 Technical assumptions

Before the proof is provided, we present and explain more technical assumptions of the paper in this section. Most of these assumptions are extended from Lam and Souza (2015a).

R1. The column vectors $\text{vec}(\boldsymbol{W}_{0i}^T)$ in $\boldsymbol{V}_0$ are linearly independent to each other, such that there exists a constant $u > 0$ with $\sigma_M^2(\boldsymbol{V}_0) \geq u > 0$ uniformly as $N \to \infty$, where $\sigma_i(A)$ is the ith largest singular value of a matrix A. Moreover, $\max_{1\leq i\leq M} \|\boldsymbol{W}_{0i}\|_1 \leq c < 1$ uniformly as $N \to \infty$ for some constant $c > 0$.

R2. Write $\boldsymbol{\epsilon}_t = \boldsymbol{\Sigma}_\epsilon^{1/2}\boldsymbol{\epsilon}_t^*$, where $\boldsymbol{\Sigma}_\epsilon$ is the covariance matrix for $\boldsymbol{\epsilon}_t$. Then the elements in $\boldsymbol{\Sigma}_\epsilon$ are all less than $\sigma_{\max}^2$ uniformly as $N \to \infty$. Same for the variance of the elements in $\boldsymbol{B}_t$. We also assume $\|\boldsymbol{\Sigma}_\epsilon^{1/2}\|_\infty \leq S_\epsilon < \infty$ uniformly as $N \to \infty$, with $\{\epsilon_{t,j}^*\}_{1\leq j\leq N}$ being a martingale difference with respect to the filtration generated by $\sigma(\epsilon_{t,1}^*, ..., \epsilon_{t,j}^*)$. The tail condition $P(|Z| > v) \leq D_1 \exp(-D_2 v^q)$ is also satisfied by $\epsilon_{t,j}^*$.

R3. All singular values of $\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)$ are uniformly larger than Nu for some constant $u > 0$, while the maximum singular value is also of order N. Individual entries in the matrix $\mathbb{E}(\boldsymbol{x}_t\boldsymbol{b}_t^T)$ are uniformly bounded away from infinity.

R4. For the same constant a, we have for each N

$$\max_{1\leq i\leq N}\sum_{j=1}^{N}\Big\|\mathbb{E}(\sum_{q\geq 0}\boldsymbol{b}_{t,i}\boldsymbol{x}_{t-q,j}^T)\Big\|_{\max}, \quad \max_{1\leq j\leq N}\sum_{i=1}^{N}\Big\|\mathbb{E}(\sum_{q\geq 0}\boldsymbol{b}_{t,i}\boldsymbol{x}_{t-q,j}^T)\Big\|_{\max} \leq C_{bx}N^a$$

where $C_{bx} > 0$ is a constant and $\boldsymbol{b}_{t,i}$, $\boldsymbol{x}_{t,j}$ are the column vectors for the ith row of $\boldsymbol{B}_t$ and jth row of $\boldsymbol{X}_t$ respectively. At the same time, assume also that $\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t\boldsymbol{\zeta})$ has all singular values of order N^{1+a}.

R5. Assume $0 < b < 1$. For fixed $= 1, \ldots, K$, the eigenvalues of $N^{-b}\text{var}(\boldsymbol{B}_{t,k})$ and $\text{var}(\boldsymbol{\epsilon}_T)$ are uniformly bounded away from 0 and infinity, and respectively dominates the singular values of the sum of $N^{-b}\text{cov}(\boldsymbol{B}_{t+\tau,k}, \boldsymbol{B}_{t,k})$ over $\tau \neq 0$ and the sum of $\mathbb{E}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T)$ over $\tau \neq 0$. Also, for each $i = 1, \ldots, N$, we assume that

$$\sum_\tau \sigma_i\big(N^{-b}\text{cov}(\boldsymbol{B}_{t+\tau,k}, \boldsymbol{B}_{t,k})\big) < \infty, \quad \sum_\tau \sigma_i\big(\mathbb{E}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau})\big) < \infty.$$

R6. Define $\lambda_T = cT^{-1/2}\log^{1/2}(T \vee N)$ for some constant $c > 0$. The tuning parameter γ_T is such that $\gamma_T = C\lambda_T$ for some constant $C > 0$.

R7. In all the assumptions above, we assume that as $N, T \to \infty$, $\lambda_T N^{1-a} = o(1)$, $N^{-a+b-1/w}\log^{-1}(T \vee N) = o(1)$, $\log(T \vee N)N^{1/w-b} = o(1)$ and $N^{b-a} = o(T\lambda_T)$.

Assumption R1 essentially requires that each specification $\boldsymbol{W}_{0i}$ is different from one another to a certain extent. This is intuitive, since if $\boldsymbol{W}_{0i}$ and $\boldsymbol{W}_{0l}$ are too similar to each other, the coefficients δ_{ji} and δ_{jl} are not well defined, and this will have a negative impact on the performance of our estimators.

The assumptions on $\boldsymbol{\Sigma}_\epsilon$ in R2 is mainly for the convenience of proofs, while the martingale difference assumption for $\boldsymbol{\epsilon}_t$ is a relaxation to independence.

Assumptions R3 and R4 are closely related. They paint a picture of how the exogenous variables in $\boldsymbol{B}_t$ are correlated with $\boldsymbol{X}_{t-q}$. Assumption R3 essentially says that the covariance between a variable in $\boldsymbol{B}_t$ and one in $\boldsymbol{X}_t$ is finite uniformly as $N \to \infty$. Then for $k = 1, \ldots, K$, considering the kth diagonal entry of $\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)$ is $\sum_{j=1}^N \mathbb{E}(X_{t,jk}B_{t,jk})$ with each $\mathbb{E}(X_{t,jk}B_{t,jk})$ being finite, it is indeed reasonable to assume that each diagonal entry in the matrix is of order N. This assumption is needed for the estimator $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\delta})$ to be well-defined.

Assumption R4 essentially describes how each row of variables in $\boldsymbol{B}_t$ are correlated with different rows of variables in $\boldsymbol{X}_t$. With this, we can actually derive easily that $\|\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t\boldsymbol{\zeta})\|_1$ has order at most N^{1+a}. Hence the assumption of having all the singular values of $\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t\boldsymbol{\zeta})$ of order N^{1+a} is reasonable.

Assumption R5 assumes a rate for the singular values of $\text{var}(\boldsymbol{B}_{t,k})$ essentially, which is important in certain asymptotic normality results. The rate N^b, possibly differing from N^a, is reasonable as well since the way that $\boldsymbol{B}_t$ and $\boldsymbol{X}_t$ are correlated do not directly indicate how the variables in $\boldsymbol{B}_t$ itself are correlated, unless of course when $\boldsymbol{B}_t = \boldsymbol{X}_t$ where $\boldsymbol{X}_t$ itself is exogenous, in which case $a = b$. The variance-covariance matrix being dominating the lag τ auto-covariances is for the ease of presentation of rates of convergence in the asymptotic normality results in this chapter.

1.8.2 Proof of theorems

The followings are Lemma 1 and 2 of Lam and Souza (2015a) respectively.

Lemma 1. *For a zero mean time series process $\mathbf{x}_t = \boldsymbol{f}(\mathcal{F})$ with dependence measure $\theta^x_{t,d,j}$ defined in Section 1.3, assume $\Theta^x_{m,a} \leq Cm^{-\alpha}$ as in Assumption M3. Then there exists constants C_1, C_2 and C_3 independent of v, T and the index j such that*

$$P(|1/T\sum_{t=1}^{T} x_{t,j}| > v) \leq \frac{C_1 T^{w(1/2-\tilde{\alpha})}}{(Tv)^w} + C_2 \exp(-C_3 T^{\tilde{\beta}} v^2),$$

where $\tilde{\alpha} = \alpha \wedge (1/2 - 1/w)$, *and* $\tilde{\beta} = (3 + 2\tilde{\alpha}w)/(1 + w)$.

Furthermore, assume another zero mean time series process z_t *(can be the same process* x_t*) with both* $\Theta^x_{m,2w}$, $\Theta^z_{m,2w} \leq Cm^{-\alpha}$, *as in Assumption M3. Then provided* $\max_j \|x_{tj}\|_{2w}$, $\max_j \|z_{tj}\|_{2w} \leq c_0 \leq \infty$ *where* c_0 *is a constant, the above Nagaev-type inequality holds for the product process* $\{x_{tj}z_{tl} - \mathbb{E}(x_{tj}z_{tl})\}$.

Lemma 2. *For any* $N \times N$ *matrix* $\boldsymbol{H} = (h_1, \ldots, h_N)^T$ *and any* $N \times K$ *matrix* $\boldsymbol{M}$, *define*

$$\boldsymbol{V_H} = \begin{pmatrix} \boldsymbol{I}_K \otimes h_1 \\ \vdots \\ \boldsymbol{I}_K \otimes h_N \end{pmatrix}.$$

Then we have

$$\boldsymbol{HM} = (\boldsymbol{I}_N \otimes \mathrm{vec}^T(\boldsymbol{M}))\boldsymbol{V_H}.$$

We first present an Theorem 5 which states that a set $\mathcal{M}$ is such that $P(\mathcal{M}) \to 1$ as $T, N \to \infty$, and our estimators enjoy nice properties on $\mathcal{M}$. This theorem is in fact exactly the same as Theorem S.1 of Lam and Souza (2015a).

Denote $B_{t,ij}$ and $X_{t,ij}$ the (i,j) entry of $\boldsymbol{B}_t$ and $\boldsymbol{X}_t$ respectively, and define $\mathcal{M} = \cap_{i=1}^{7}\mathcal{A}_i$, where

$$\mathcal{A}_1 = \left\{ \max_{1\leq i,k\leq N} \max_{1\leq j,l\leq K} |T^{-1}\sum_{t=1}^{T}[B_{t,ij}X_{t,kl} - \mathbb{E}(B_{t,ij}X_{t,kl})]| < \lambda_T \right\},$$

$$\mathcal{A}_2 = \left\{ \max_{1\leq i,k\leq N} \max_{1\leq j\leq K} |T^{-1}\sum_{t=1}^{T} B_{t,ij}\epsilon_{t,k}| < \lambda_T \right\},$$

$$\mathcal{A}_3 = \left\{ \max_{1\leq k\leq K} |T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{N} B_{t,sk}\epsilon_{t,s}| < \lambda_T N^{1/2+1/2w} \right\},$$

$$\mathcal{A}_4 = \left\{ \max_{1\leq i\leq N} \max_{1\leq j\leq K} |\bar{B}_{\cdot,ij} - \mathbb{E}(B_{t,ij})| < \lambda_T \right\},$$

$$\mathcal{A}_5 = \left\{ \max_{1\leq j\leq N} |\bar{\epsilon}_{\cdot,j}| < \lambda_T \right\},$$

$$\mathcal{A}_6 = \left\{ \max_{1\leq i\leq N} \max_{1\leq j\leq K} |\bar{X}_{\cdot,ij}| < \lambda_T \right\},$$

$$\mathcal{A}_7 = \left\{ \max_{1\leq k\leq K} |\sum_{s=1}^{N} \bar{B}_{\cdot,sk}\bar{\epsilon}_{\cdot,s}| < 2^{1/2}\lambda_T N^{1/2}\log^{1/2}(T \vee N) S_\epsilon(\max_{i.j} |\mathbb{E}(B_{t,ij})| + \lambda_T) \right\}.$$

Theorem 5. *Let Assumptions M1-M4 in Section 1.3.1 and R1-R7 in Section 1.8.1 hold. Suppose $\alpha \geq 1/2 - 1/w$ in Assumption M3, and for the application of the Nagaev-type inequality in Lemma 1 for the processes defined in $\mathcal{A}_1$ to $\mathcal{A}_7$, suppose the constants C_1, C_2 and C_3 are the same. Then with $c \geq \sqrt{3/C_3}$ where c is the constant defined in $\lambda_T = cT^{-1/2}\log^{1/2}(T \vee N)$, we have*

$$P(\mathcal{M}) \geq 1 - 8C_1K^2(C_3/3)^{w/2}\frac{N^2}{T^{w/2-1}\log^{w/2}(T \vee N)} - \frac{8C_2K^2N^2}{T^3 \vee N^3} - \frac{2K}{T \vee N}.$$

It approaches 1 if we assume further that $N = o(T^{w/4-1/2}\log^{w/4}(T))$.

Proof of Theorem 1

From (1.4) and that

$$\begin{aligned} y_0^v &= \sum_{i=1}^{M} \delta_{0i} \boldsymbol{W}_{0i}^{\otimes} y_0^v + \sum_{j=1}^{p} \Big(\sum_{i=1}^{M} \delta_{ji} \boldsymbol{W}_{0i}^{\otimes} \Big) y_j^v + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^v + \mathbf{1}_T \otimes \boldsymbol{\mu} \\ &= \Big(\boldsymbol{I}_{TN} - \sum_{i=1}^{M} \delta_{0i} \boldsymbol{W}_{0i}^{\otimes} \Big)^{-1} \Big(\sum_{j=1}^{p} \Big(\sum_{i=1}^{M} \delta_{ji} \boldsymbol{W}_{0i}^{\otimes} \Big) y_j^v + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^v + \mathbf{1}_T \otimes \boldsymbol{\mu} \Big), \end{aligned}$$

it is easy to get, since $\boldsymbol{B}^{vT}(\mathbf{1}_T \otimes \boldsymbol{\mu}) = \mathbf{0}$, that

$$\boldsymbol{\beta}(\boldsymbol{\delta}) - \boldsymbol{\beta} = (\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{\epsilon}^v.$$

Moreover,

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \boldsymbol{\beta}(\hat{\boldsymbol{\delta}}) \\ &= (\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \Big[\Big(\boldsymbol{I}_{TN} - \sum_{i=1}^{M} \hat{\delta}_{0i} \boldsymbol{W}_{0i}^{\otimes} \Big) y_0^v - \sum_{j=1}^{p} \Big(\sum_{i=1}^{M} \hat{\delta}_{ji} \boldsymbol{W}_{0i}^{\otimes} \Big) y_j^v \Big] \\ &= (\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \Big[\Big(\boldsymbol{I}_{TN} - \sum_{i=1}^{M} \delta_{0i} \boldsymbol{W}_{0i}^{\otimes} \Big) y_0^v - \sum_{j=1}^{p} \Big(\sum_{i=1}^{M} \delta_{ji} \boldsymbol{W}_{0i}^{\otimes} \Big) y_j^v \\ &+ \sum_{i=1}^{M} (\delta_{0i} - \hat{\delta}_{0i}) \boldsymbol{W}_{0i}^{\otimes} y_0^v + \sum_{j=1}^{p} \Big(\sum_{i=1}^{M} (\delta_{ji} - \hat{\delta}_{ji}) \boldsymbol{W}_{0i}^{\otimes} \Big) y_j^v \Big] \\ &= \boldsymbol{\beta}(\boldsymbol{\delta}) + (\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \\ &\qquad \cdot \Big[\sum_{i=1}^{M} (\delta_{0i} - \hat{\delta}_{0i}) \boldsymbol{W}_{0i}^{\otimes} y_0^v + \sum_{j=1}^{p} \Big(\sum_{i=1}^{M} (\delta_{ji} - \hat{\delta}_{ji}) \boldsymbol{W}_{0i}^{\otimes} \Big) y_j^v \Big]. \end{aligned}$$

Using the above, we can decompose

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = I_0 + I_1 + I_2 + I_3 + I_4 + I_5, \text{ where}$$

$$I_0 = (\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t))^{-1}(\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t) - T^{-2}\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

$$I_1 = (\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t))^{-1} T^{-2}\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} \boldsymbol{\epsilon}^v,$$

$$I_2 = (\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t))^{-1} T^{-2}\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} (\sum_{i=1}^{M} (\delta_{0i} - \hat{\delta}_{0i}) \boldsymbol{W}_{0i}^{\otimes}) \boldsymbol{\Pi}^{\otimes} \boldsymbol{X}\boldsymbol{\beta},$$

$$I_3 = (\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t))^{-1} T^{-2}\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} (\sum_{i=1}^{M} (\delta_{0i} - \hat{\delta}_{0i}) \boldsymbol{W}_{0i}^{\otimes}) \boldsymbol{\Pi}^{\otimes} \boldsymbol{\epsilon}^v,$$

$$I_4 = (\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t))^{-1} T^{-2}\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} (\sum_{i=1}^{M} (\delta_{0i} - \hat{\delta}_{0i}) \boldsymbol{W}_{0i}^{\otimes}) \boldsymbol{\Pi}^{\otimes} (\sum_{j=1}^{p} (\sum_{i=1}^{M} \delta_{ji} \boldsymbol{W}_{0i}^{\otimes}) y_j^v),$$

$$I_5 = (\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t))^{-1} T^{-2}\boldsymbol{X}^T \boldsymbol{B}^v \boldsymbol{B}^{vT} (\sum_{j=1}^{p} (\sum_{i=1}^{M} (\delta_{ji} - \hat{\delta}_{ji}) \boldsymbol{W}_{0i}^{\otimes} y_j^v)),$$

with $\boldsymbol{\Pi}^{\otimes} = (\boldsymbol{I}_{TN} - \sum_{i=1}^{M} \delta_{0i} \boldsymbol{W}_{0i}^{\otimes})^{-1}$. We need to find the rate of convergence of I_0 I_1, I_2, I_3, I_4 and I_5.

To this end, using Assumption R3 in Section 1.8.1,

$$\|\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t)^{-1}\|_1 \leq \frac{K^{1/2}}{\lambda_{\min}(\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T \boldsymbol{X}_t))} \leq \frac{K^{1/2}}{N^2 u^2}.$$

Define $\boldsymbol{U} = \boldsymbol{I}_N \otimes T^{-1} \sum_{t=1}^{T} \text{vec}(\boldsymbol{B}_t - \bar{\boldsymbol{B}}) \text{vec}^T(\boldsymbol{X}_t)$ and $\boldsymbol{U}_0 = \boldsymbol{I}_N \otimes \mathbb{E}(\boldsymbol{b}_t \boldsymbol{x}_t^T)$, then we can write $\mathrm{T}^{-1}\boldsymbol{X}^T \boldsymbol{B}^v = \boldsymbol{V}_{\boldsymbol{I}_N}^T \boldsymbol{U} \boldsymbol{V}_{\boldsymbol{I}_N}$ and $\mathbb{E}(\boldsymbol{X}_t^T \boldsymbol{B}_t) = \boldsymbol{V}_{\boldsymbol{I}_N}^T \boldsymbol{U}_0 \boldsymbol{V}_{\boldsymbol{I}_N}$. Also, denote $\boldsymbol{W}_j^c$,$\boldsymbol{B}_{t,j}$ and $\boldsymbol{X}_{t,j}$ the jth column of $\boldsymbol{W}$, $\boldsymbol{B}_t$ and $\boldsymbol{X}_t$ respectively, and let $\boldsymbol{\pi}_j^T$ be the

jth row of $\mathbf{\Pi}$. Then on $\mathcal{M}$,

$$
\begin{aligned}
&\|I_0\|_1 \\
&\leq \|(\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}\|_1\|((\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t)) - T^{-2}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_1 \\
&\leq \frac{K^{1/2}}{N^2u^2}\Big[\|\boldsymbol{V}_{\boldsymbol{I}_N}^T(\boldsymbol{U}_0-\boldsymbol{U})^T\boldsymbol{V}_{\boldsymbol{I}_N}\boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}_0\|_1 \\
&+ \|\boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}^T\boldsymbol{V}_{\boldsymbol{I}_N}\boldsymbol{V}_{\boldsymbol{I}_N}^T(\boldsymbol{U}_0-\boldsymbol{U})\|_1\Big]\|\boldsymbol{V}_{\boldsymbol{I}_N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_1 \\
&\leq \frac{K^{1/2}}{N^2u^2}\Big[K\|\boldsymbol{U}_0-\boldsymbol{U}\|_{\max}\cdot N\cdot K\|\boldsymbol{U}_0\|_{\max} \\
&+ (K\|\boldsymbol{V}_{\boldsymbol{I}_N}^T(\boldsymbol{U}-\boldsymbol{U}_0)^T\boldsymbol{V}_{\boldsymbol{I}_N}\|_{\max} + K\|\boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}_0^T\boldsymbol{V}_{\boldsymbol{I}_N}\|_{\max})\cdot K\|\boldsymbol{U}_0-\boldsymbol{U}\|_{\max}\Big]\cdot N\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \\
&\leq K^{1/2}(2\lambda_T\sigma_{bx}(1+\mu_{b,\max}+\lambda_T)+\lambda_T^2(1+\mu_{b,\max}+\lambda_T)^2)\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \\
&= O(\lambda_T\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1),
\end{aligned}
$$

where $\mu_{b,\max} = \big\|\mathbb{E}(\boldsymbol{b}_t)\big\|_{\max}$. At the same time, on $\mathcal{M}$,

$$
\begin{aligned}
\|I_1\|_1 &\leq \frac{K^{1/2}}{N^2u^2}\|T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\|_1\|T^{-1}\boldsymbol{B}^{vT}\boldsymbol{\epsilon}^v\|_1 \\
&\leq \frac{K^{1/2}}{N^2u^2}\|\boldsymbol{V}_{\boldsymbol{I}_N}^T(\boldsymbol{U}-\boldsymbol{U}_0)\boldsymbol{V}_{\boldsymbol{I}_N} + \boldsymbol{V}_{\boldsymbol{I}_N}^T\boldsymbol{U}_0^T\boldsymbol{V}_{\boldsymbol{I}_N}\|_1 \\
&\quad\cdot(K\lambda_TN^{1/2+1/2w} + \sqrt{2}K\lambda_TN^{1/2}\log(T\vee N)S_\epsilon(\mu_{b,max}+\lambda_T)) \\
&\leq \frac{K^{1/2}}{N^2u^2}N(\lambda_T(1+\mu_{b,\max}+\lambda_T)+\sigma_{bx}) \\
&\quad\cdot(K\lambda_TN^{1/2+1/2w} + \sqrt{2}K\lambda_TN^{1/2}\log(T\vee N)S_\epsilon(\mu_{b,\max}+\lambda_T)) \\
&= O(\lambda_TN^{-1/2+1/2w}).
\end{aligned}
$$

Recall that $\boldsymbol{W}_j = \sum_{i=1}^M \delta_{ji}\boldsymbol{W}_{0i}$, and denoting $\widehat{\boldsymbol{W}}_j = \sum_{i=1}^M \widehat{\delta}_{ji}\boldsymbol{W}_{0i}$ for $j = 0, 1, \ldots, p$,

then on $\mathcal{M}$,

$$\begin{aligned}
\|I_2\|_1 &\leq \frac{K^{1/2}}{N^2u^2}\|T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\|_1\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T(\boldsymbol{W}_0-\hat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\boldsymbol{X}_t\|_1\|\boldsymbol{\beta}\|_1 \\
&\leq \frac{K^{1/2}}{N^2u^2}O(N)\Big(K\cdot\max_{1\leq r\leq K}\Big|\sum_{j=1}^{N}(\boldsymbol{W}_{0,j}^c-\hat{\boldsymbol{W}}_{0,j}^c)^TT^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_{t,r}-\bar{\boldsymbol{B}}_{.,r})\boldsymbol{X}_{t,r}^T\boldsymbol{\pi}_j\Big|\Big) \\
&\leq O(N^{-1})(\sum_{j=1}^{N}(\lambda_T(1+\mu_{b,\max}+\lambda_T)+\sigma_{bx})\|\boldsymbol{W}_{0,j}^c-\hat{\boldsymbol{W}}_{0,j}^c\|_1\|\boldsymbol{\pi}_j\|_1) \\
&\leq O(N^{-1})(N\|\boldsymbol{\delta}_0-\hat{\boldsymbol{\delta}}_0\|_1)=O(\|\boldsymbol{\delta}_0-\hat{\boldsymbol{\delta}}_0\|_1).
\end{aligned}$$

Similarly, on $\mathcal{M}$,

$$\begin{aligned}
\|I_3\|_1 &\leq \frac{K^{1/2}}{N^2u^2}\|T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\|_1\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T(\boldsymbol{W}_0-\hat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\boldsymbol{\epsilon}_t\|_1 \\
&\leq \frac{K^{1/2}}{N^2u^2}O(N)\Big(K\max_{1\leq r\leq K}\Big|\sum_{j=1}^{N}(\boldsymbol{W}_{0,j}^c-\hat{\boldsymbol{W}}_{0,j}^c)^T\Big(T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_{t,r}-\bar{\boldsymbol{B}}_{.,r})\boldsymbol{\epsilon}_t^T\Big)\boldsymbol{\pi}_j\Big|\Big) \\
&\leq O(N^{-1})\cdot O(N\lambda_T\max_{1\leq j\leq N}\big\|\boldsymbol{\pi}_j\big\|_1\max_{1\leq j\leq N}\big\|\boldsymbol{W}_{0,j}^c-\widehat{\boldsymbol{W}}_{0,j}^c\big\|_1)=O(\lambda_T\big\|\boldsymbol{\delta}_0-\widehat{\boldsymbol{\delta}}_0\big\|_1).
\end{aligned}$$

For bounding $\big\|I_4\big\|_1$ and $\big\|I_5\big\|_1$, recall that from Section 1.3.1, we can express y_t as

$$y_t=\boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}(\boldsymbol{\mu}+\boldsymbol{X}_t\boldsymbol{\beta}+\boldsymbol{\epsilon}_t)=\sum_{q\geq 0}\boldsymbol{\Psi}_q\boldsymbol{\Pi}(\boldsymbol{\mu}+\boldsymbol{X}_{t-q}\boldsymbol{\beta}+\boldsymbol{\epsilon}_{t-q}),\tag{1.9}$$

where $\boldsymbol{\Psi}_q$ is $N\times N$ such that $\sum_{q\geq 0}\big\|\boldsymbol{\Psi}_q\big\|_\infty<\infty$ because of stationarity. Then we can decompose

$$\begin{aligned}
\|I_4\|_1 &\leq \frac{K^{1/2}}{N^2u^2}\|T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\|_1 \\
&\quad\cdot\Big\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T(\boldsymbol{W}_0-\hat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\sum_{j=1}^{p}\boldsymbol{W}_j\boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}(\boldsymbol{X}_{t-j}\boldsymbol{\beta}+\boldsymbol{\epsilon}_{t-j})\Big\|_1 \\
&= O(N^{-1}(\|I_{41}\|_1+\|I_{42}\|_1)),\text{ where} \\
\|I_{41}\|_1 &= \Big\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T(\boldsymbol{W}_0-\hat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\sum_{j=1}^{p}\boldsymbol{W}_j\boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}\boldsymbol{X}_{t-j}\boldsymbol{\beta}\Big\|_1, \\
\big\|I_{42}\big\|_1 &= \Big\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T(\boldsymbol{W}_0-\hat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\sum_{j=1}^{p}\boldsymbol{W}_j\boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}\boldsymbol{\epsilon}_{t-j}\Big\|_1.
\end{aligned}$$

On $\mathcal{M}$, we have

$$\begin{aligned}\big\|I_{41}\big\|_1 &\le \max_{1\le j\le p}\max_{1\le r,k\le K} pK^2\big\|\boldsymbol{\beta}\big\|_1 \\ &\quad\cdot\left|\sum_{q\ge 0}\Big\{T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_{t,r}-\bar{\boldsymbol{B}}_{\cdot,r})^T(\boldsymbol{W}_0-\widehat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\boldsymbol{W}_j\boldsymbol{\Psi}_q\boldsymbol{\Pi}\boldsymbol{X}_{t-q-j,k}\Big\}\right| \\ &= O(N\sigma_{bx}\big\|\boldsymbol{W}_0-\widehat{\boldsymbol{W}}_0\big\|_\infty\big\|\boldsymbol{W}_j\big\|_\infty\big\|\boldsymbol{\Pi}\big\|_\infty^2\sum_{q\ge 0}\big\|\boldsymbol{\Psi}_q\big\|_\infty) \\ &= O(N\big\|\boldsymbol{\delta}_0-\widehat{\boldsymbol{\delta}}_0\big\|_1),\end{aligned}$$

where the second line is by Assumption R4. At the same time on $\mathcal{M}$,

$$\begin{aligned}&\big\|I_{42}\big\|_1 \\ &\le \max_{1\le j\le p}\max_{1\le r\le K} pK\left|\sum_{q\ge 0}\Big\{T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_{t,r}-\bar{\boldsymbol{B}}_{\cdot,r})^T(\boldsymbol{W}_0-\widehat{\boldsymbol{W}}_0)\boldsymbol{\Pi}\boldsymbol{W}_j\boldsymbol{\Psi}_q\boldsymbol{\Pi}\boldsymbol{\epsilon}_{t-q-j}\Big\}\right| \\ &= O(N\lambda_T\big\|\boldsymbol{W}_0-\widehat{\boldsymbol{W}}_0\big\|_\infty\big\|\boldsymbol{W}_j\big\|_\infty\big\|\boldsymbol{\Pi}\big\|_\infty^2\sum_{q\ge 0}\big\|\boldsymbol{\Psi}_q\big\|_\infty) \\ &= O(N\lambda_T\big\|\boldsymbol{\delta}_0-\widehat{\boldsymbol{\delta}}_0\big\|_1),\end{aligned}$$

where the second line follows from the rate on $\mathcal{A}_2$. These imply that on $\mathcal{M}$,

$$\big\|I_4\big\|_1 = O(\big\|\boldsymbol{\delta}_0-\widehat{\boldsymbol{\delta}}_0\big\|_1).$$

To bound $\big\|I_5\big\|_1$, we can decompose

$$\begin{aligned}&\|I_5\|_1 \\ &\le \frac{K^{1/2}}{N^2u^2}\|T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v\|_1\Big\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T\sum_{j=1}^{p}(\boldsymbol{W}_j-\hat{\boldsymbol{W}}_j)\boldsymbol{\Phi}^{-1}(L)\boldsymbol{\Pi}(\boldsymbol{X}_{t-j}\boldsymbol{\beta}+\boldsymbol{\epsilon}_{t-j})\Big\|_1 \\ &= O(N^{-1}(\|I_{51}\|_1+\|I_{52}\|_1)),\end{aligned}$$

where

$$\|I_{51}\|_1 = \Big\| T^{-1} \sum_{t=1}^{T} (\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T \sum_{j=1}^{p} (\boldsymbol{W}_j - \hat{\boldsymbol{W}}_j) \boldsymbol{\Phi}^{-1}(L) \boldsymbol{\Pi} \boldsymbol{X}_{t-j} \boldsymbol{\beta} \Big\|_1,$$
$$\|I_{52}\|_1 = \Big\| T^{-1} \sum_{t=1}^{T} (\boldsymbol{B}_t - \bar{\boldsymbol{B}})^T \sum_{j=1}^{p} (\boldsymbol{W}_j - \hat{\boldsymbol{W}}_j) \boldsymbol{\Phi}^{-1}(L) \boldsymbol{\Pi} \boldsymbol{\epsilon}_{t-j} \Big\|_1.$$

To bound $\big\|I_{51}\big\|_1$, similar to the treatment on $\big\|I_{41}\big\|_1$, on $\mathcal{M}$,

$$\begin{aligned}
\big\|I_{51}\big\|_1 &\leq \max_{1\leq r,k\leq K} K^2 \big\|\boldsymbol{\beta}\big\|_1 \Big| T^{-1} \sum_{t=1}^{T} (\boldsymbol{B}_{t,r} - \bar{\boldsymbol{B}}_{\cdot,r})^T \sum_{j=1}^{p} (\boldsymbol{W}_j - \widehat{\boldsymbol{W}}_j) \sum_{q\geq 0} \boldsymbol{\Psi}_q \boldsymbol{\Pi} \boldsymbol{X}_{t-q-j,k} \Big| \\
&= O(N\sigma_{bx} \sum_{j=1}^{p} \big\|\boldsymbol{W}_j - \widehat{\boldsymbol{W}}_j\big\|_\infty \big\|\boldsymbol{\Pi}\big\|_\infty \sum_{q\geq 0} \big\|\boldsymbol{\Psi}_q\big\|_\infty) \\
&= O(N\big\|\boldsymbol{\delta} - \widehat{\boldsymbol{\delta}}\big\|_1).
\end{aligned}$$

Finally, on $\mathcal{M}$,

$$\begin{aligned}
\big\|I_{52}\big\|_1 &\leq \max_{1\leq r\leq K} K \Big| T^{-1} \sum_{t=1}^{T} (\boldsymbol{B}_{t,r} - \bar{\boldsymbol{B}}_{\cdot,r})^T \sum_{j=1}^{p} (\boldsymbol{W}_j - \widehat{\boldsymbol{W}}_j) \sum_{q\geq 0} \boldsymbol{\Psi}_q \boldsymbol{\Pi} \boldsymbol{\epsilon}_{t-q-j} \Big| \\
&= O(N\lambda_T \sum_{j=1}^{p} \big\|\boldsymbol{W}_j - \widehat{\boldsymbol{W}}_j\big\|_\infty \big\|\boldsymbol{\Pi}\big\|_\infty \sum_{q\geq 0} \big\|\boldsymbol{\Psi}_q\big\|_\infty) \\
&= O(N\lambda_T \big\|\boldsymbol{\delta} - \widehat{\boldsymbol{\delta}}\big\|_1).
\end{aligned}$$

Hence on $\mathcal{M}$, we have

$$\big\|I_5\big\|_1 = O(\big\|\boldsymbol{\delta} - \widehat{\boldsymbol{\delta}}\big\|_1).$$

Combining the rates for $\big\|I_0\big\|_1$ to $\big\|I_5\big\|_1$, we can conclude that on $\mathcal{M}$,

$$\big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big\|_1 = O(\lambda_T N^{-1/2+1/2w} + \big\|\boldsymbol{\delta} - \widehat{\boldsymbol{\delta}}\big\|_1). \tag{1.10}$$

We need to find the order of $\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1$. From (1.3) and (1.5), it is easy to show that

$$\boldsymbol{K} y_0^v - \boldsymbol{B}^T y = \boldsymbol{K} y_0^v - (\boldsymbol{B}^T \boldsymbol{\epsilon} + \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V} \boldsymbol{\delta} + \boldsymbol{B}^T \boldsymbol{X}_{\boldsymbol{\beta}} \mathrm{vec}(\boldsymbol{I}_N)),$$

where

$$\begin{aligned}\boldsymbol{K}y_0^v &= \boldsymbol{H}\boldsymbol{\delta} + \boldsymbol{K}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{K}\boldsymbol{\epsilon}^v \\ &= \boldsymbol{H}\boldsymbol{\delta} + T^{-1/2}N^{-a/2}\sum_{t=1}^{T}\boldsymbol{X}_t \otimes (\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\zeta}\boldsymbol{\beta} + \boldsymbol{K}\boldsymbol{\epsilon}^v \\ &= \boldsymbol{H}\boldsymbol{\delta} + \boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}}\mathrm{vec}(\boldsymbol{I}_N) + \boldsymbol{K}\boldsymbol{\epsilon}^v.\end{aligned}$$

Hence,

$$\boldsymbol{K}y_0^v - \boldsymbol{B}^T y = -\boldsymbol{B}^T\boldsymbol{\epsilon} + (\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\boldsymbol{\delta} + \boldsymbol{K}\boldsymbol{\epsilon}^v.$$

Substituting the above back to (1.5), we can decompose

$$\begin{aligned}\hat{\boldsymbol{\delta}} - \boldsymbol{\delta} &= \left[(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\right]^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T\left[\boldsymbol{K}\boldsymbol{\epsilon}^v - \boldsymbol{B}^T\boldsymbol{\epsilon}\right] \\ &= D_1 + D_2, \quad \text{where} \\ D_1 &= \left[(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\right]^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T\boldsymbol{K}\boldsymbol{\epsilon}^v, \\ D_2 &= -\left[(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\right]^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T\boldsymbol{B}^T\boldsymbol{\epsilon}.\end{aligned}$$

To bound $\left\|D_1\right\|_1$ and $\left\|D_2\right\|_1$, we introduce some notations and find their L_1 norm bounds first. For $i = 1, \ldots, M$, define

$$\boldsymbol{U}_q = \boldsymbol{I}_N \otimes T^{-1}\sum_{t=1}^{T}\mathrm{vec}(\boldsymbol{B}_t - \bar{\boldsymbol{B}})\mathrm{vec}^T(\boldsymbol{X}_{t-q}), \quad \boldsymbol{U}_{0q} = \boldsymbol{I}_N \otimes \mathbb{E}(\boldsymbol{b}_t\boldsymbol{x}_{t-q}^T).$$

Also, define for $i = 1, \ldots, M$ and $j = 1, \ldots, p$,

$$\begin{aligned}
&\boldsymbol{A}_1 = T^{-1} \sum_{t=1}^{T} \boldsymbol{X}_t \otimes (\boldsymbol{B}_t - \bar{\boldsymbol{B}})\boldsymbol{\zeta}, \qquad \boldsymbol{A}_1^0 = \mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t\boldsymbol{\zeta}), \\
&\boldsymbol{A}_2 = (\boldsymbol{V}_{\boldsymbol{I}_N}^T \boldsymbol{U}^T \boldsymbol{V}_{\boldsymbol{I}_N} \boldsymbol{V}_{\boldsymbol{I}_N}^T \boldsymbol{U} \boldsymbol{V}_{\boldsymbol{I}_N})^{-1}, \quad \boldsymbol{A}_2^0 = (\boldsymbol{V}_{\boldsymbol{I}_N}^T \boldsymbol{U}_0^T \boldsymbol{V}_{\boldsymbol{I}_N} \boldsymbol{V}_{\boldsymbol{I}_N}^T \boldsymbol{U}_0 \boldsymbol{V}_{\boldsymbol{I}_N})^{-1}, \\
&\boldsymbol{A}_3 = \boldsymbol{V}_{\boldsymbol{I}_N}^T \boldsymbol{U}^T \boldsymbol{V}_{\boldsymbol{I}_N}, \qquad \boldsymbol{A}_3^0 = \boldsymbol{V}_{\boldsymbol{I}_N}^T \boldsymbol{U}_0^T \boldsymbol{V}_{\boldsymbol{I}_N}, \\
&\boldsymbol{A}_{4ij} = \sum_{q=0}^{\infty} \boldsymbol{V}_{\boldsymbol{W}_{0i}^T}^T \boldsymbol{U}_{q+j} \boldsymbol{V}_{\tilde{\boldsymbol{\Pi}}_q} \boldsymbol{\beta}, \qquad \boldsymbol{A}_{4ij}^0 = \sum_{q=0}^{\infty} \boldsymbol{V}_{\boldsymbol{W}_{0i}^T}^T \boldsymbol{U}_{0,q+j} \boldsymbol{V}_{\tilde{\boldsymbol{\Pi}}_q} \boldsymbol{\beta}, \\
&\boldsymbol{A}_{5ij} = \sum_{q=0}^{\infty} \boldsymbol{V}_{\boldsymbol{W}_{0i}^T}^T \Big(\boldsymbol{I}_N \otimes T^{-1} \sum_{t=1}^{T} \mathrm{vec}(\boldsymbol{B}_t - \bar{\boldsymbol{B}}) \boldsymbol{\epsilon}_{t-q-j}^T \Big) \mathrm{vec}(\tilde{\boldsymbol{\Pi}}_q^T).
\end{aligned} \tag{1.11}$$

where $\tilde{\boldsymbol{\Pi}}_q = \boldsymbol{\Psi}_q \boldsymbol{\Pi}$. It is straightforward to see that, on $\mathcal{M}$,

$$\|\boldsymbol{A}_1 - \boldsymbol{A}_1^0\|_{\max} = O(\lambda_T) \tag{1.12}$$

Meanwhile, by Assumptions R4 and R7, on $\mathcal{M}$,

$$\|\boldsymbol{A}_1\|_1 \leq \|\boldsymbol{A}_1^0\|_1 + \|\boldsymbol{A}_1 - \boldsymbol{A}_1^0\|_1 = O(N^{1+a} + \lambda_T N^2) = O(N^{1+a}). \tag{1.13}$$

Similarly, on $\mathcal{M}$,

$$\|\boldsymbol{A}_3^0\|_1 \leq K \|\boldsymbol{V}_{\boldsymbol{I}_N}^T \boldsymbol{U}_0^T \boldsymbol{V}_{\boldsymbol{I}_N}\|_{\max} = O(N), \quad \|\boldsymbol{A}_3 - \boldsymbol{A}_3^0\|_1 = O(\lambda_T N), \quad \|\boldsymbol{A}_3\|_1 = O(N). \tag{1.14}$$

As $\boldsymbol{A}_2^0 = (\boldsymbol{A}_3^0 \boldsymbol{A}_3^{0T})^{-1}$,

$$\|\boldsymbol{A}_2^0\|_1 \leq \frac{K^{1/2}}{\lambda_{\min}(\boldsymbol{A}_3^0 \boldsymbol{A}_3^{0T})} \leq \frac{K^{1/2}}{N^2 u^2} = O(N^{-2}). \tag{1.15}$$

Moreover, we know that $\boldsymbol{A}_2 - \boldsymbol{A}_2^0 = (\boldsymbol{A}_2 - \boldsymbol{A}_2^0)((\boldsymbol{A}_2^0)^{-1} - \boldsymbol{A}_2^{-1})\boldsymbol{A}_2^0 + \boldsymbol{A}_2^0((\boldsymbol{A}_2^0)^{-1} - \boldsymbol{A}_2^{-1})\boldsymbol{A}_2^0$, and on $\mathcal{M}$,

$$\begin{aligned}
&\|(\boldsymbol{A}_2^0)^{-1} - \boldsymbol{A}_2^{-1}\|_1 \\
&= \|\boldsymbol{A}_3^0 \boldsymbol{A}_3^{0T} - \boldsymbol{A}_3 \boldsymbol{A}_3^T\|_1 \leq \|\boldsymbol{A}_3^0 - \boldsymbol{A}_3\|_1 \|\boldsymbol{A}_3^{0T}\|_1 + \|\boldsymbol{A}_3\|_1 \|\boldsymbol{A}_3^{0T} - \boldsymbol{A}_3^T\|_1 = O(\lambda_T N^2).
\end{aligned}$$

Therefore, on $\mathcal{M}$,

$$\|\boldsymbol{A}_2 - \boldsymbol{A}_2^0\|_1 \leq \frac{\|(\boldsymbol{A}_2^0)^{-1} - \boldsymbol{A}_2^{-1}\|_1 \|\boldsymbol{A}_2^0\|_1^2}{1 - O(\lambda_T N^2 N^{-2})} = O\Big(\frac{\lambda_T N^2 N^{-4}}{1 - \lambda_T N^2 N^{-2}}\Big) = O(\lambda_T N^{-2}). \quad (1.16)$$

As for $\|\boldsymbol{A}_{4ij}\|_1$, by Assumptions M1 and R4, defining $\tilde{\boldsymbol{\pi}}_{q,r}^T$ to be the rth row of $\tilde{\boldsymbol{\Pi}}_q$, we have on $\mathcal{M}$,

$$\|\boldsymbol{A}_{4ij}^0\|_1 \leq \sum_{q=0}^{\infty} K\|\boldsymbol{\beta}\|_1 \|\boldsymbol{V}_{\boldsymbol{W}_{0i}^T}^T \boldsymbol{U}_{0,q+j} \boldsymbol{V}_{\tilde{\boldsymbol{\Pi}}_q}\|_{\max} \quad (1.17)$$

$$= \sum_{q=0}^{\infty} K\|\boldsymbol{\beta}\|_1 \max_{1\leq k,m\leq K} \Big| \sum_{r=1}^{N} \boldsymbol{W}_{0i,r}^{cT} \mathbb{E}(\boldsymbol{X}_{t-q-j,k} \boldsymbol{B}_{t,m}^T) \tilde{\boldsymbol{\pi}}_{q,r} \Big|$$

$$= O(\|\boldsymbol{W}_{0i}\|_1 \sum_{q\geq 0} \|\tilde{\boldsymbol{\Pi}}_q\|_\infty \cdot N) \leq O(\|\boldsymbol{W}_{0i}\|_1 \sum_{q\geq 0} (\|\boldsymbol{\Pi}\|_\infty \|\boldsymbol{\Psi}_q\|_\infty) \cdot N) = O(N). \quad (1.18)$$

Similarly, we can easily show on $\mathcal{M}$ that

$$\|\boldsymbol{A}_{4ij} - \boldsymbol{A}_{4ij}^0\|_1 = O(\lambda_T N).$$

Hence we have

$$\|\boldsymbol{A}_{4ij}\|_1 = O(N). \quad (1.19)$$

To bound $\|\boldsymbol{A}_{5ij}\|_1$, an element in $\boldsymbol{A}_{5ij}$ is bounded on $\mathcal{M}$ by

$$\Big| \sum_{r=1}^{N} \sum_{q=0}^{\infty} \boldsymbol{W}_{0i,r}^{cT} T^{-1} \sum_{t=1}^{T} (\boldsymbol{B}_{t,k} - \bar{\boldsymbol{B}}_k) \boldsymbol{\epsilon}_{t-q-j}^T \tilde{\boldsymbol{\pi}}_{q,r} \Big| = O(\lambda_T N), \quad \text{so} \quad \|\boldsymbol{A}_{5ij}\|_1 = O(\lambda_T N). \quad (1.20)$$

We now decompose $D_1 = F_1 + F_2 + F_3$, where

$$
\begin{aligned}
F_1 &= \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})\right]^{-1} \\
&\quad \cdot \Big[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) - T^{-1}N^a(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\Big] D_1, \\
F_2 &= \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})\right]^{-1} \\
&\quad \cdot (T^{-1/2}N^{a/2}\boldsymbol{H} - \boldsymbol{H}_{20} - T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V} + \boldsymbol{H}_{10})^T \cdot T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v, \\
F_3 &= \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})\right]^{-1} (\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T \cdot T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v.
\end{aligned}
$$

Both $\boldsymbol{H}_{20}$ and $\boldsymbol{H}_{10}$ are $N^2 \times M(p+1)$ matrices defined in Theorem 3. By Assumptions R3 and R4, it is easy to show that

$$
\begin{aligned}
\sigma_M(\boldsymbol{H}_{20}) &\geq \sigma_K(\boldsymbol{A}_1^0)\sigma_K(\boldsymbol{A}_2^0)\sigma_K(\boldsymbol{A}_3^0)\sigma_{\min}(\boldsymbol{A}_{410}^0, \ldots, \boldsymbol{A}_{4M0}^0, \ldots, \boldsymbol{A}_{41p}^0, \ldots, \boldsymbol{A}_{4Mp}^0) \\
&\geq \frac{CN^{1+a} \cdot N \cdot N}{\lambda_{\max}(\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))} \geq CN^{1+a}, \\
\sigma_M^2(\boldsymbol{H}_{10}) &\geq \sigma_M^2(\boldsymbol{V}_0)\sigma_N^2\Big((\boldsymbol{I}_N \otimes \boldsymbol{\zeta}_T)\sum_{q=0}^{\infty}\mathbb{E}(\mathrm{vec}(\boldsymbol{B}_t^T)\mathrm{vec}(\boldsymbol{X}_t^T)^T)(\boldsymbol{I}_N \otimes \boldsymbol{\beta})\tilde{\boldsymbol{\Pi}}_q\Big) \geq CN^{1+a}.
\end{aligned}
$$

Hence the smallest singular value of $\boldsymbol{H}_{20}$ dominates that of $\boldsymbol{H}_{10}$, and so for some constant $u > 0$,

$$
\sigma_{M(p+1)}^2(\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) \geq uN^{1+a}. \tag{1.21}
$$

With this, we have

$$
\|[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})]^{-1}\|_1 \leq \frac{M^{1/2}(p+1)^{1/2}}{\lambda_{\min}[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})]} \leq \frac{M^{1/2}(p+1)^{1/2}}{uN^{1+a}} \tag{1.22}
$$

To bound $\|D_1\|_1$, using (1.22), we have

$$\|F_1\|_1 \leq \frac{M^{3/2}(p+1)^{1/2}}{N^{1+a}u}\Big[\|\boldsymbol{H}_{20}-\boldsymbol{H}_{10}\|_1 \tag{1.23}$$
$$\cdot\Big(\|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_{\max}+\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}-\boldsymbol{H}_{10}\|_{\max}\Big)$$
$$+\|T^{-1/2}N^{a/2}(\boldsymbol{H}-\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\|_{\max}$$
$$\cdot\Big(\|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_1+\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}-\boldsymbol{H}_{10}\|_1\Big)\Big]\|D_1\|_1, \tag{1.24}$$
$$\|F_2\|_1 \leq \frac{M^{3/2}(p+1)^{1/2}}{N^{1+a}\mu}\Big(\|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_{\max}+\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}-\boldsymbol{H}_{10}\|_{\max}\Big)$$
$$\cdot\|T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v\|_1, \tag{1.25}$$
$$\|F_3\|_1 \leq \frac{M^{3/2}(p+1)^{1/2}}{N^{1+a}\mu}\|\boldsymbol{H}_{20}-\boldsymbol{H}_{10}\|_{\max}\cdot\|T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v\|_1. \tag{1.26}$$

Now, to bound $\|F_1\|_1$, $\|F_2\|_1$ and $\|F_3\|_1$, we consider

$$\|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_{\max}=\max_{1\leq i\leq M}\max_{1\leq j\leq p}\|\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3(\boldsymbol{A}_{4ij}+\boldsymbol{A}_{5ij})-\boldsymbol{A}_1^0\boldsymbol{A}_2^0\boldsymbol{A}_3^0\boldsymbol{A}_{4ij}^0\|_{\max}$$
$$\leq \max_{1\leq i\leq M}\max_{1\leq j\leq p}\|\boldsymbol{A}_1\|_{\max}\|\boldsymbol{A}_2\|_1\|\boldsymbol{A}_3\|_1\|\boldsymbol{A}_{5ij}\|_1+$$
$$\max_{1\leq i\leq M}\max_{1\leq j\leq p}\Big[\|\boldsymbol{A}_1\|_{\max}\|\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{A}_{4ij}-\boldsymbol{A}_2^0\boldsymbol{A}_3^0\boldsymbol{A}_{4ij}^0\|_1+\|\boldsymbol{A}_1-\boldsymbol{A}_1^0\|_{\max}\|\boldsymbol{A}_2^0\boldsymbol{A}_3^0\boldsymbol{A}_{4ij}^0\|_1\Big], \tag{1.27}$$

with

$$\|\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{A}_{4ij}-\boldsymbol{A}_2^0\boldsymbol{A}_3^0\boldsymbol{A}_{4ij}^0\|_1\leq\max_{1\leq j\leq p}\Big[\|\boldsymbol{A}_2\|_1\|\boldsymbol{A}_3-\boldsymbol{A}_3^0\|_1\|\boldsymbol{A}_{4ij}\|_1+$$
$$\|\boldsymbol{A}_2\|_1\|\boldsymbol{A}_3^0\|_1\|\boldsymbol{A}_{4ij}-\boldsymbol{A}_{4ij}^0\|_1+\|\boldsymbol{A}_2-\boldsymbol{A}_2^0\|_1\|\boldsymbol{A}_3^0\|_1\|\boldsymbol{A}_{4ij}^0\|_1\Big].$$

Therefore, base on the rates found in (1.12) to (1.20), and (1.27), we have on $\mathcal{M}$ that

$$\|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_{\max}=O(\lambda_T),\quad\text{and}\quad\|T^{-1/2}N^{a/2}\boldsymbol{H}-\boldsymbol{H}_{20}\|_1=O(\lambda_T N^2). \tag{1.28}$$

Define $\mathbf{L}^q=T^{-1}\sum_{t=1}^T\text{vec}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T)\text{vec}^T(\boldsymbol{X}_{t-q})$ and $\mathbf{L}_0^q=\mathbb{E}(\text{vec}(\boldsymbol{B}_t^T)\text{vec}^T(\boldsymbol{X}_{t-q}^T)$.

Then, by Assumption R4 and on $\mathcal{M}$, we have

$$
\begin{aligned}
&\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\|_1 \\
&= \max_{0\le l\le p}\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\zeta}y_{t-l}\|_1 = \|T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\zeta}y_t^T\|_1 \\
&\le \|T^{-1}\sum_{t=1}^{T}\sum_{q=0}^{\infty}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\zeta}(\boldsymbol{\Psi}_q\boldsymbol{\Pi}(\boldsymbol{X}_{t-q}\boldsymbol{\beta}+\boldsymbol{\epsilon}_{t-q}))^T\|_1 \\
&\le O(\lambda_T N + N^a + \lambda_T N) = O(N^a), \quad \text{and} \qquad (1.29)\\
&\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}-\boldsymbol{H}_{10}\|_{\max} \\
&= \max_{0\le l\le p}\max_{1\le i\le M,1\le j\le N}\|\Big(T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\zeta}y_{t-l}^T-\mathbb{E}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\zeta}y_{t-l}^T)\Big)\boldsymbol{W}_{0i,j}^c\|_{\max} \\
&= \max_{1\le i\le M,1\le j\le N}\|\Big(T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\zeta}y_{t}^T-\mathbb{E}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})\boldsymbol{\zeta}y_{t}^T)\Big)\boldsymbol{W}_{0i,j}^c\|_{\max} \\
&\le \max_{1\le i\le M,1\le j\le N}\Big[\|\boldsymbol{I}_N\otimes\boldsymbol{\zeta}^T\|_\infty\sum_{q\ge 0}\|\mathbf{L}^q-\mathbf{L}_0^q\|_{\max}\|\tilde{\boldsymbol{\Pi}}_q^T\|_1\|\boldsymbol{I}_N\otimes\boldsymbol{\beta}\|_1\|\boldsymbol{W}_{0i,j}^c\|_1 \\
&+\|\boldsymbol{I}_N\otimes\boldsymbol{\zeta}^T\|_\infty\sum_{q\ge 0}\|T^{-1}\sum_{t=1}^{T}\mathrm{vec}((\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T)\boldsymbol{\epsilon}_{t-q}^T\|_{\max}\|\tilde{\boldsymbol{\Pi}}_q^T\|_1\|\boldsymbol{W}_{0i,j}^c\|_1\Big] \\
&= O(\lambda_T). \qquad (1.30)
\end{aligned}
$$

Hence on $\mathcal{M}$,

$$\|T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}-\boldsymbol{H}_{10}\|_1 = O(\lambda_T N^2). \qquad (1.31)$$

Using the rates found in (1.12) to (1.20), on $\mathcal{M}$ (in particular using the rate on $\mathcal{A}_3$),

$$\|T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v\|_1 = \Big\|\boldsymbol{A}_1\boldsymbol{A}_2\boldsymbol{A}_3(T^{-1}\sum_{t=1}^{T}(\boldsymbol{B}_t-\bar{\boldsymbol{B}})^T\epsilon_t)\Big\|_1 = O(\lambda_T N^{1/2+1/2w+a}). \qquad (1.32)$$

Therefore, using results from (1.22) to (1.32), we know that

$$\begin{aligned}\|D_1\|_1 &\leq \frac{M^{3/2}}{N^{1+a}u}\Big(o(\lambda_T N^2) + o(1)o(\lambda_T N^2 + \lambda_T N^2)\Big)\|D_1\|_1 \\ &+ \frac{M^{3/2}}{N^{1+a}u}\Big(O(\lambda_T)O(\lambda_T N^{1/2+1/2w+a})\Big) + \frac{M^{3/2}}{N^{1+a}u}O(\lambda_T N^{1/2+1/2w+a}) \\ &= O(\lambda_T N^{-1/2+1/2w}).\end{aligned}$$

For the rate of $\big\|D_2\big\|_1$, we refer to the proof of asymptotic normality of $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}$ in Theorem 3 for the proof of the asymptotic normality of D_2 (along the exact same lines of proofs as in Theorem 3). Therefore, we state here the result that

$$T^{1/2}(\boldsymbol{M}_2\mathbf{S}_2\boldsymbol{M}_2^T)^{-1/2}D_2 \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{I}_m),$$

where $\mathbf{S}_2$ is defined in Theorem 3, and $\boldsymbol{M}_2 = \big[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})\big]^{-1}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T$. By Assumption R5, we conclude that all the eigenvalues of $\mathbf{S}_2$ are of order N^b. Hence by (1.21),

$$\begin{aligned}\lambda_{\max}(\boldsymbol{M}_2\mathbf{S}_2\boldsymbol{M}_2^T) &\leq \lambda_{\max}(\mathbf{S}_2)\lambda_{\max}(\big[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})\big]^{-1}) \\ &\leq \frac{\lambda_{\max}(\mathbf{S}_2)}{\sigma^2_{M(p+1)}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})} = O(N^{-1-a+b}),\end{aligned}$$

which can also be derived as the order for the lower bound of $\lambda_{\min}(\boldsymbol{M}_2\mathbf{S}_2\boldsymbol{M}_2^T)$. Hence we have $\|D_2\|_1 = O_p(T^{-1/2}N^{-(1+a-b)/2})$.

Finally, by Assumption R7 and the result of Theorem 5,

$$\begin{aligned}\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1 &= O_P(\big\|D_1\big\|_1 + \big\|D_2\big\|_1) = O_P(\lambda_T \cdot N^{1/2+a+1/2w}) + O_P(T^{-1/2}N^{-(1+a-b)/2}) \\ &= O_P(\lambda_T \cdot N^{-1/2+1/2w}).\end{aligned}$$

At the same time, using the result above,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 = O_p(\lambda_T N^{-1/2+1/2w} + \|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\|_1) = O_p(\lambda_T N^{-1/2+1/2w}). \quad \square$$

Proof of Theorem 2.

It has been shown that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \sum_{i=0}^5 I_i$ in the proof of Theorem 1. From the rate

of $\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1$ and Assumption R7, it is clear that

$$\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_1 = O_P(\lambda_T N^{-1/2+1/2w}) = o_P(T^{-1/2}N^{-(1-b)/2}).$$

Therefore, if we can prove that I_1 is $T^{1/2}N^{(1-b)/2}$-convergent, then I_1 dominates I_2 to I_5, while $\left\|I_0\right\|_1 = O_P(\lambda_T\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_1) = o_P(\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_1)$.

We now prove that for $\boldsymbol{\alpha} \in \mathbb{R}^K$ such that $\left\|\boldsymbol{\alpha}\right\| = 1$, $\boldsymbol{\alpha}^T I_1$ is $T^{1/2}N^{(1-b)/2}$-convergent by proving its asymptotic normality. Recall that

$$\begin{aligned}
I_1 &= (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}T^{-2}\boldsymbol{X}^T\boldsymbol{B}^v\boldsymbol{B}^{vT}\boldsymbol{\epsilon}^v \\
&= (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}(T^{-1}\boldsymbol{X}^T\boldsymbol{B}^v - \mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t))T^{-1}\boldsymbol{B}^{vT}\boldsymbol{\epsilon}^v \\
&\quad + (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)T^{-1}\boldsymbol{B}^{vT}\boldsymbol{\epsilon}^v.
\end{aligned}$$

It is easy to show that the second term above dominates the first. Therefore, if we can prove that

$$\sum_{t\geq 0}\|\boldsymbol{P}_0(\boldsymbol{\alpha}^T\boldsymbol{M}_1\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t)\| < \infty, \tag{1.33}$$

where $\boldsymbol{M}_1 = (\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{X}_t))^{-1}\mathbb{E}(\boldsymbol{X}_t^T\boldsymbol{B}_t)$, by Theorem 3(ii) of Wu (2011), we then have

$$T^{1/2}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{\alpha})^{-1/2}\boldsymbol{\alpha}^T I_1 \xrightarrow{\mathcal{D}} \boldsymbol{N}(0,1),$$

where $\boldsymbol{\Sigma}_1 = \boldsymbol{M}_1\sum_{\tau\in\mathbb{Z}}\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T\boldsymbol{B}_{t+\tau})\boldsymbol{M}_1^T$.

To determine the rate of the eigenvalues in $\boldsymbol{\Sigma}_1$, consider the (k,k) element of $\sum_\tau\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T\boldsymbol{B}_{t+\tau})$,

$$\begin{aligned}
&\sum_\tau\mathbb{E}(\boldsymbol{B}_{t,k}^T\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T\boldsymbol{B}_{t+\tau,k}) = \sum_\tau\mathrm{tr}(\mathbb{E}(\boldsymbol{B}_{t+\tau,k}\boldsymbol{B}_{t,k}^T)\mathbb{E}(\boldsymbol{\epsilon}_{t+\tau}\boldsymbol{\epsilon}_t^T)) \\
&= \sum_\tau\mathrm{tr}(\mathrm{cov}(\boldsymbol{B}_{t+\tau,k}\boldsymbol{B}_{t,k})\mathrm{cov}(\boldsymbol{\epsilon}_{t+\tau}\boldsymbol{\epsilon}_t)) + \sum_\tau\mathrm{tr}(\boldsymbol{\mu}_{b,k}\boldsymbol{\mu}_{b,k}^T\mathrm{cov}(\boldsymbol{\epsilon}_{t+\tau}\boldsymbol{\epsilon}_t)).
\end{aligned}$$

By Assumptions R5, the first term is N^{1+b}-convergent exactly and the second term's rate is

$$\sum_\tau\boldsymbol{\mu}_{b,k}^T\mathrm{cov}(\boldsymbol{\epsilon}_{t+\tau}\boldsymbol{\epsilon}_t)\boldsymbol{\mu}_{b,k} \leq \lambda_{\max}(\sum_\tau\mathrm{cov}(\boldsymbol{\epsilon}_{t+\tau}\boldsymbol{\epsilon}_t))\|\boldsymbol{\mu}_{b,k}\|^2 = O(\|\boldsymbol{\mu}_{b,k}\|^2) = O(N).$$

Since K is finite, the order of the eigenvalues of $\sum_\tau\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T\boldsymbol{B}_{t+\tau})$ is exactly

N^{1+b}. Also, for $i = 1, \ldots, K$,

$$\lambda_{\min}(\boldsymbol{M}_1\boldsymbol{M}_1^T)\lambda_{\min}(\sum_{\tau}\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T\boldsymbol{B}_{t+\tau}))$$
$$\leq \lambda_i(\boldsymbol{\Sigma}_1) \leq \lambda_{\max}(\boldsymbol{M}_1\boldsymbol{M}_1^T)\lambda_{\max}(\sum_{\tau}\mathbb{E}(\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^T\boldsymbol{B}_{t+\tau})).$$

Since the order of the eigenvalues of $\boldsymbol{M}_1\boldsymbol{M}_1^T$ is N^{-2}, the order of all the eigenvalues of $\boldsymbol{\Sigma}_1$ is exactly N^{-1+b}. It means also that $\boldsymbol{\alpha}^T I_1$ is indeed $T^{1/2}N^{(1-b)/2}$-convergent, and so I_1 is $T^{1/2}N^{(1-b)/2}$-convergent in particular since K is finite. With the asymptotic normality for $\boldsymbol{\alpha}^T I_1$, we can then use the multivariate version of Theorem 3(ii) of Wu (2011) to conclude that

$$T^{1/2}\boldsymbol{\Sigma}_1^{-1/2}I_1 \xrightarrow{\mathcal{D}} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}_K),$$

where we replaced $\boldsymbol{\alpha}$ by I_K.

It remains to prove (1.33). We decompose

$$\boldsymbol{P}_0(\boldsymbol{\alpha}^T\boldsymbol{M}_1\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t) = \boldsymbol{\alpha}^T\boldsymbol{M}_1\boldsymbol{P}_0(\boldsymbol{B}_t^T)\mathbb{E}_0(\boldsymbol{\epsilon}_t) + \boldsymbol{\alpha}^T\boldsymbol{M}_1\mathbb{E}_{-1}(\boldsymbol{B}_t^T)\boldsymbol{P}_0(\boldsymbol{\epsilon}_t),$$

so that we have $\|\boldsymbol{P}_0(\boldsymbol{\alpha}^T\boldsymbol{M}_1\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t)\| \leq C_{1,t} + C_{2,t}$, where

$$\begin{aligned}
C_{1,t}^2 &= \mathbb{E}(\boldsymbol{\alpha}^T\boldsymbol{M}_1\boldsymbol{P}_0(\boldsymbol{B}_t^T)\mathbb{E}_0(\boldsymbol{\epsilon}_t)\mathbb{E}_0(\boldsymbol{\epsilon}_t^T)\boldsymbol{P}_0(\boldsymbol{B}_t)\boldsymbol{M}_1^T\boldsymbol{\alpha}) \\
&\leq \boldsymbol{\alpha}^T\boldsymbol{M}_1\mathbb{E}(\boldsymbol{P}_0(\boldsymbol{B}_t^T)\boldsymbol{P}_0(\boldsymbol{B}_t))\boldsymbol{M}_1\boldsymbol{\alpha}\mathbb{E}(\lambda_{\max}(\mathbb{E}_0(\boldsymbol{\epsilon}_t)\mathbb{E}_0(\boldsymbol{\epsilon}_t^T))) \\
&\leq \left\|\boldsymbol{\alpha}^T\boldsymbol{M}_1\right\|^2\lambda_{\max}(\mathbb{E}(\boldsymbol{P}_0(\boldsymbol{B}_t^T)\boldsymbol{P}_0(\boldsymbol{B}_t)))\mathbb{E}(\mathbb{E}_0(\boldsymbol{\epsilon}_t^T)\mathbb{E}_0(\boldsymbol{\epsilon}_t)) \\
&= O(N^{-1}\max_{1\leq k\leq K}\mathbb{E}(\boldsymbol{P}_0(\boldsymbol{B}_{t,k}^T)\boldsymbol{P}_0(\boldsymbol{B}_{t,k}))\mathbb{E}(N^{-1}\mathbb{E}_0(\boldsymbol{\epsilon}_t^T)\mathbb{E}_0(\boldsymbol{\epsilon}_t))) \\
&= O(\max_{1\leq k\leq K}\max_{1\leq s\leq N}\|\boldsymbol{P}_0(B_{t,sk})\|^2\max_{1\leq j\leq N}\mathbb{E}(\mathbb{E}_0^2(\epsilon_{t,j}))) \\
&= O(\max_{1\leq k\leq K}\max_{1\leq s\leq N}\|\boldsymbol{P}_0(B_{t,sk})\|^2\sigma_{\max}^2), \qquad (1.34)
\end{aligned}$$

so that $\sum_{t\geq 0} C_{1,t} < \infty$ by our assumption $\sum_{t\geq 0}\max_{1\leq k\leq K}\max_{1\leq s\leq N}\|\boldsymbol{P}_0^b(B_{t,sk})\| < \infty$.

Similarly, we have

$$
\begin{aligned}
C_{2,t}^2 &= \mathbb{E}(\boldsymbol{\alpha}^T \boldsymbol{M}_1 \mathbb{E}_{-1}(\boldsymbol{B}_t^T) \boldsymbol{P}_0(\boldsymbol{\epsilon}_t) \boldsymbol{P}_0(\boldsymbol{\epsilon}_t^T) \mathbb{E}_{-1}(\boldsymbol{B}_t) \boldsymbol{M}_1^T \boldsymbol{\alpha}) \\
&\leq \boldsymbol{\alpha}^T \boldsymbol{M}_1 \mathbb{E}(\mathbb{E}_{-1}(\boldsymbol{B}_t^T) \mathbb{E}_{-1}(\boldsymbol{B}_t)) \boldsymbol{M}_1^T \boldsymbol{\alpha} \mathbb{E}(\lambda_{\max}(\boldsymbol{P}_0(\boldsymbol{\epsilon}_t) \boldsymbol{P}_0(\boldsymbol{\epsilon}_t^T))) \\
&\leq \left\| \boldsymbol{\alpha}^T \boldsymbol{M}_1 \right\|^2 \lambda_{\max}(\mathbb{E}(\mathbb{E}_{-1}(\boldsymbol{B}_t^T) \mathbb{E}_{-1}(\boldsymbol{B}_t))) \mathbb{E}(\boldsymbol{P}_0(\boldsymbol{\epsilon}_t^T) \boldsymbol{P}_0(\boldsymbol{\epsilon}_t)) \\
&= O(\max_{1 \leq k \leq K} \max_{1 \leq s \leq N} \mathbb{E}(\mathbb{E}_{-1}^2(B_{t,sk})) \max_{1 \leq j \leq N} \|\boldsymbol{P}_0(\epsilon_{t,j})\|^2) \\
&= O((\sigma_{\max}^2 + \max_{s,k} \mu_{b,sk}^2) \max_{1 \leq j \leq N} \|\boldsymbol{P}_0(\epsilon_{t,j})\|^2) \\
&= O(\max_{1 \leq j \leq N} \|\boldsymbol{P}_0^{\epsilon}(\epsilon_{t,j})\|^2), \qquad (1.35)
\end{aligned}
$$

so that $\sum_{t \geq 0} C_{2,t} < \infty$ by our assumption of $\sum_{t \geq 0} \max_{1 \leq j \leq N} \|\boldsymbol{P}_0^{\epsilon}(\epsilon_{t,j})\| < \infty$. Hence (1.33) is established, and the proof of the theorem is completed. □

Proof of Theorem 3.

To prove the asymptotic normality of $\hat{\boldsymbol{\delta}}$, we need to apply the same method we used for the proof of Theorem 2. Recall that from the proof of Theorem 1,

$$
\begin{aligned}
&\hat{\boldsymbol{\delta}} - \boldsymbol{\delta} = \left[(\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}) \right]^{-1} (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T \left[\boldsymbol{K} \boldsymbol{\epsilon}^v - \boldsymbol{B}^T \boldsymbol{\epsilon} \right], \\
&\hat{\boldsymbol{\delta}} - \boldsymbol{\delta} = D_1 + D_2, \quad \text{where} \\
&\quad D_1 = \left[(\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}) \right]^{-1} (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T \boldsymbol{K} \boldsymbol{\epsilon}^v, \\
&\quad D_2 = - \left[(\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}) \right]^{-1} (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T \boldsymbol{B}^T \boldsymbol{\epsilon}.
\end{aligned}
$$

Moreover, we further decompose D_1 as in the proof of Theorem 1 such that $D_1 = F_1 + F_2 + F_3$, where

$$
\begin{aligned}
F_1 &= \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T (\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) \right]^{-1} \\
&\quad \cdot \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T (\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) - T^{-1} N^a (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V})^T (\boldsymbol{H} - \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V}) \right] D_1, \\
F_2 &= \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T (\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) \right]^{-1} \\
&\quad \cdot (T^{-1/2} N^{a/2} \boldsymbol{H} - \boldsymbol{H}_{20} - T^{-1/2} N^{a/2} \boldsymbol{B}^T \boldsymbol{Z} \boldsymbol{V} + \boldsymbol{H}_{10})^T \cdot T^{-1/2} N^{a/2} \boldsymbol{K} \boldsymbol{\epsilon}^v, \\
F_3 &= \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T (\boldsymbol{H}_{20} - \boldsymbol{H}_{10}) \right]^{-1} (\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T \cdot T^{-1/2} N^{a/2} \boldsymbol{K} \boldsymbol{\epsilon}^v. \qquad (1.36)
\end{aligned}
$$

From the proof of Theorem 1, it is clear that F_3 dominates all other terms in the

decomposition of D_1. As for D_2, we can apply similar decomposition, and the term

$$F_4 = -\left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})\right]^{-1}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T T^{-1/2} N^{a/2} \boldsymbol{B}^T \boldsymbol{\epsilon} \tag{1.37}$$

dominates in the decomposition of D_2. Hence to show the asymptotic normality of $\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}$, we only consider

$$\begin{aligned} F_3 + F_4 &= T^{-1/2} N^{a/2} \boldsymbol{M}_2(\boldsymbol{K}\boldsymbol{\epsilon}^v - \boldsymbol{B}^T\boldsymbol{\epsilon}) \\ &= T^{-1}\sum_{t=1}^{T} \boldsymbol{M}_2(\boldsymbol{M}\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t - \mathrm{vec}(\boldsymbol{B}_t\boldsymbol{\zeta}\boldsymbol{\epsilon}_t^T))(1 + o_P(1)), \end{aligned}$$

where $\boldsymbol{M}_2 = \left[(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})\right]^{-1}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})^T$. In view of the above and Theorem 3(ii) of Wu (2011), to prove the asymptotic normality of $\boldsymbol{\alpha}^T(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})$ where $\boldsymbol{\alpha} \in \mathbb{R}^{M(p+1)}$, if we can show that

$$\sum_{t\geq 0} \left\| \boldsymbol{P}_0(\boldsymbol{\alpha}^T\boldsymbol{M}_2(\boldsymbol{M}\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t - \mathrm{vec}(\boldsymbol{B}_t\boldsymbol{\zeta}\boldsymbol{\epsilon}_t^T)))\right\| < \infty, \tag{1.38}$$

then we can conclude by Theorem 3(ii) of Wu (2011) that

$$T^{1/2}(\boldsymbol{\alpha}^T\boldsymbol{\Sigma}_2\boldsymbol{\alpha})^{-1/2}\boldsymbol{\alpha}^T(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}) \xrightarrow{\mathcal{D}} \boldsymbol{N}(0,1), \tag{1.39}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_2 &= \sum_{\tau\in\mathbb{Z}} \boldsymbol{M}_2 \mathrm{cov}(\boldsymbol{M}\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t - \mathrm{vec}(\boldsymbol{B}_t\boldsymbol{\zeta}\boldsymbol{\epsilon}_t^T), \boldsymbol{M}\boldsymbol{B}_{t+\tau}^T\boldsymbol{\epsilon}_{t+\tau} - \mathrm{vec}(\boldsymbol{B}_{t+\tau}\boldsymbol{\zeta}\boldsymbol{\epsilon}_{t+\tau}^T))\boldsymbol{M}_2^T \\ &= \boldsymbol{M}_2(\mathbf{S}_1 + \mathbf{S}_2 - \mathbf{S}_3 - \mathbf{S}_3^T)\boldsymbol{M}_2^T, \end{aligned}$$

with $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$ as defined in the statement of the theorem. A generalization to Theorem 3(ii) of Wu (2011) then gives us the asymptotic normality result after replacing $\boldsymbol{\alpha}$ by $\boldsymbol{I}_{M(p+1)}$.

It remains to show (1.38). Consider

$$
\begin{aligned}
&\left\|\boldsymbol{\alpha}^T \boldsymbol{M}_2 \boldsymbol{M}\right\|^2 \\
&\leq \lambda_{\max}(\boldsymbol{M}_2 \boldsymbol{M}_2^T)\lambda_{\max}(\boldsymbol{M}\boldsymbol{M}^T) \\
&= O(N^{-1-a}) \cdot O(N^{-2}) \cdot \lambda_{\max}(\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t\boldsymbol{\zeta})\mathbb{E}(\boldsymbol{X}_t^T \otimes \boldsymbol{\zeta}^T\boldsymbol{B}_t^T)) \\
&= O(N^{-3-a}) \cdot \left\|\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{B}_t\boldsymbol{\zeta})\right\|_1 \left\|\mathbb{E}(\boldsymbol{X}_t^T \otimes \boldsymbol{\zeta}^T\boldsymbol{B}_t^T)\right\|_1 \\
&= O(N^{-3-a}) \cdot O(N^{1+a}) \cdot O(1) = O(N^{-2}),
\end{aligned}
$$

where the last line follows from Assumption R4. Then similar to showing (1.34), by the above, we have

$$
\left\|\boldsymbol{P}_0(\boldsymbol{\alpha}^T \boldsymbol{M}_2 \boldsymbol{M}\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t)\right\| = O(\max_{1\leq k\leq K}\max_{1\leq s\leq N}\left\|\boldsymbol{P}_0(B_{t,sk})\right\|), \tag{1.40}
$$

so that $\sum_{t\geq 0}\left\|P_0(\boldsymbol{\alpha}^T \boldsymbol{M}_2 \boldsymbol{M}\boldsymbol{B}_t^T\boldsymbol{\epsilon}_t)\right\| < \infty$ by the assumptions of the theorem. At the same time,

$$
\begin{aligned}
\boldsymbol{P}_0(\boldsymbol{\alpha}^T \boldsymbol{M}_2 \mathrm{vec}(\boldsymbol{B}_t\boldsymbol{\zeta}\boldsymbol{\epsilon}_t^T)) &= \boldsymbol{P}_0(\boldsymbol{\alpha}^T \boldsymbol{M}_2(\boldsymbol{\epsilon}_t \otimes \boldsymbol{B}_t\boldsymbol{\zeta})) \\
&= \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{M}_2\Big(\mathbb{E}_0(\boldsymbol{\epsilon}_t) \otimes \mathbb{E}_0(\boldsymbol{B}_t\boldsymbol{\zeta}) - \mathbb{E}_{-1}(\boldsymbol{\epsilon}_t) \otimes \mathbb{E}_{-1}(\boldsymbol{B}_t\boldsymbol{\zeta})\Big) \\
&= \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{M}_2 \boldsymbol{P}_0(\boldsymbol{\epsilon}_t) \otimes \mathbb{E}_0(\boldsymbol{B}_t\gamma) + \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{M}_2 \mathbb{E}_{-1}(\boldsymbol{\epsilon}_t) \otimes \boldsymbol{P}_0(\boldsymbol{B}_t\gamma).
\end{aligned}
$$

Hence denote by $\boldsymbol{b}_{t,j}^T$ the jth row of $\boldsymbol{B}_t$,

$$
\begin{aligned}
&\big\|\boldsymbol{P}_0(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}_2\boldsymbol{\epsilon}_t \otimes \boldsymbol{B}_t\boldsymbol{\zeta})\big\| \\
&\le \Big\{2\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}_2\mathbb{E}(\boldsymbol{P}_0(\boldsymbol{\epsilon}_t)\boldsymbol{P}_0(\boldsymbol{\epsilon}_t)^{\mathrm{T}}) \otimes \mathbb{E}\big(\mathbb{E}_0(\boldsymbol{B}_t\boldsymbol{\zeta})\mathbb{E}_0(\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{B}_t^{\mathrm{T}})\big)\boldsymbol{M}_2^{\mathrm{T}}\boldsymbol{\alpha}\Big\}^{1/2} \\
&\quad + \Big\{2\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}_2\mathbb{E}(\mathbb{E}_{-1}(\boldsymbol{\epsilon}_t)\mathbb{E}_{-1}(\boldsymbol{\epsilon}_t)^{\mathrm{T}}) \otimes \mathbb{E}\big(\boldsymbol{P}_0(\boldsymbol{B}_t\boldsymbol{\zeta})\boldsymbol{P}_0(\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{B}_t^{\mathrm{T}})\big)\boldsymbol{M}_2^{\mathrm{T}}\boldsymbol{\alpha}\Big\}^{1/2} \\
&\le 2^{1/2}\big\|\boldsymbol{\alpha}\big\|_1\big\|\boldsymbol{M}_2\big\|_\infty \max_{1\le j\le N}\big\|\boldsymbol{P}_0(\epsilon_{tj})\big\| \cdot \max_{1\le j\le N}\mathrm{var}^{1/2}(\boldsymbol{b}_{t,j}^{\mathrm{T}}\boldsymbol{\zeta}) \\
&\quad + 2^{1/2}\big\|\boldsymbol{\alpha}\big\|_1\big\|\boldsymbol{M}_2\big\|_\infty \cdot \sigma_{\max} \cdot \max_{1\le j\le N}\big\|\boldsymbol{P}_0(\boldsymbol{b}_{t,j}^{\mathrm{T}}\boldsymbol{\zeta})\big\| \\
&\le 2^{1/2}\big\|\boldsymbol{\alpha}\big\|_1\big\|\boldsymbol{M}_2\big\|_\infty \max_{1\le j\le N}\big\|\boldsymbol{P}_0(\epsilon_{tj})\big\| \cdot \sigma_{\max}\big\|\boldsymbol{\zeta}\big\|_1 \\
&\quad + 2^{1/2}\big\|\boldsymbol{\alpha}\big\|_1\big\|\boldsymbol{M}_2\big\|_\infty \cdot \sigma_{\max} \cdot \max_{\substack{1\le j\le N\\1\le k\le K}}\big\|\boldsymbol{P}_0(B_{t,jk})\big\|\big\|\boldsymbol{\zeta}\big\|_1 \\
&= O(\max_{1\le j\le N}\big\|\boldsymbol{P}_0(\epsilon_{tj})\big\| + \max_{\substack{1\le j\le N\\1\le k\le K}}\big\|\boldsymbol{P}_0(B_{t,jk})\big\|),
\end{aligned}
$$

where the second inequality used the decomposition

$$\mathrm{var}(\cdot) = \mathrm{var}(\mathbb{E}_i(\cdot)) + \mathbb{E}(\mathrm{var}_i(\cdot)) \ge \mathrm{var}(\mathbb{E}_i(\cdot)),$$

and the third inequality used Assumption R2, while the last equality used $\big\|\boldsymbol{\zeta}\big\|_1 = 1$ and $\big\|\boldsymbol{M}_2\big\|_\infty = O(1)$. Hence $\sum_{t\ge 0}\big\|\boldsymbol{P}_0(\boldsymbol{\alpha}^T\boldsymbol{M}_2\mathrm{vec}(\boldsymbol{B}_t\boldsymbol{\zeta}\boldsymbol{\epsilon}_t^T)\big\| < \infty$, and together with (1.40), (1.38) is established. This completes the proof of the theorem. □

Proof of Theorem 4.

By the KKT condition, there exists a solution $\tilde{\boldsymbol{\delta}}$ to (1.7) if and only if there exists a subgradient

$$\boldsymbol{h} = \partial(\boldsymbol{u}^{\mathrm{T}}|\tilde{\boldsymbol{\delta}}|) = \Big\{\boldsymbol{h} \in \mathbb{R}^{M(p+1)} : \begin{cases} h_i = u_i\mathrm{sign}(\tilde{\delta}_i), & \tilde{\delta}_i \ne 0; \\ |h_i| \le u_i, & \text{otherwise.} \end{cases}\Big\},$$

such that differentiating the expression on the right hand side of (1.7) with respect to $\boldsymbol{\delta}$, we get

$$T^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})\tilde{\boldsymbol{\delta}} - T^{-1}(\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V} - \boldsymbol{H})^T(\boldsymbol{B}^T y - \boldsymbol{g}) = -\gamma_T\boldsymbol{h}.$$

We use a single index $i = 1, \ldots, M(p+1)$ to denote an element of $\boldsymbol{\delta}$ for easier

notation in this proof. Since we have $\boldsymbol{B}^T y = \boldsymbol{B}^T \boldsymbol{Z}\boldsymbol{V}\boldsymbol{\delta} + \boldsymbol{B}^T \boldsymbol{X}_{\boldsymbol{\beta}}\text{vec}(\boldsymbol{I}_N) + \boldsymbol{B}^T\boldsymbol{\epsilon}$, the above equation can be rewritten as

$$
\begin{aligned}
&T^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})(\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}) \\
&+ T^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T(\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}}\text{vec}(\boldsymbol{I}_N) + \boldsymbol{H}\boldsymbol{\delta} - \boldsymbol{g}) \\
&+ T^{-1}(\boldsymbol{H} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V})^T\boldsymbol{B}^T\boldsymbol{\epsilon} = -\gamma_T\boldsymbol{h}.
\end{aligned}
$$

We can show easily that $-\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}(\boldsymbol{\delta})}\text{vec}(\boldsymbol{I}_N) = \boldsymbol{H}\boldsymbol{\delta} - \boldsymbol{g}$, and hence there exists a sign consistent solution $\tilde{\boldsymbol{\delta}}$ if and only if

$$
\begin{cases}
T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)(\tilde{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H) \\
\quad +T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T(\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}-\boldsymbol{\beta}(\boldsymbol{\delta})}\text{vec}(\boldsymbol{I}_N)) \\
\quad\quad +T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T\boldsymbol{B}^T\boldsymbol{\epsilon} = -\gamma_T\boldsymbol{h}_H, \\
\quad |T^{-1}(\boldsymbol{H}_{H^c} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_{H^c})^T\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}-\boldsymbol{\beta}(\boldsymbol{\delta})}\text{vec}(\boldsymbol{I}_N) \\
\quad\quad +T^{-1}(\boldsymbol{H}_{H^c} - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_{H^c})^T\boldsymbol{B}^T\boldsymbol{\epsilon}| \leq -\gamma_T\boldsymbol{h}_{H^c},
\end{cases}
\tag{1.41}
$$

where $H = \{j : \delta_j \neq 0\}$.

From the first equation in (1.41), we decompose $\tilde{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H = I_0 + I_1 + I_2 + I_3$, where

$$
\begin{aligned}
I_0 &= -(N^{-a}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H)^{-1}(T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H) \\
&\quad - N^{-a}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H)(\tilde{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H), \\
I_1 &= (N^{-a}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H)^{-1}T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T\boldsymbol{K}\boldsymbol{\epsilon}^v, \\
I_2 &= -(N^{-a}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H)^{-1}\gamma_T\boldsymbol{h}, \\
I_3 &= -(N^{-a}(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20} - \boldsymbol{H}_{10})_H)^{-1}T^{-1}(\boldsymbol{H}_H - \boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T\boldsymbol{B}^T\boldsymbol{\epsilon}.
\end{aligned}
$$

The term I_1 has its form because of the identity $\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}(\boldsymbol{\delta})-\boldsymbol{\beta}}\text{vec}(\boldsymbol{I}_N) = \boldsymbol{K}\boldsymbol{\epsilon}^v$. Similar to bounding $\left\|F_1\right\|_1$ to $\left\|F_3\right\|_1$ in (1.24) to (1.26) in the proof of Theorem 1, we can show that

$$
\|I_0\|_{\max} = o_p(\lambda_T N^{1-a}\|\tilde{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H\|_{\max}), \quad \|I_2\|_{\max} = O(\lambda_T N^{-1}).
$$

We can show easily that

$$
\begin{aligned}
I_1 &= [(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H]^{-1}\Big(T^{-1/2}N^{a/2}(\boldsymbol{H}_H-\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T\Big)(T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v) \\
&= [(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H]^{-1}(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H^T(T^{-1/2}N^{a/2}\boldsymbol{K}\boldsymbol{\epsilon}^v)(1+o_P(1)), \\
I_3 &= -[(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H]^{-1}\Big(T^{-1/2}N^{a/2}(\boldsymbol{H}_H-\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_H)^T\Big)(T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{\epsilon}) \\
&= -[(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H]^{-1}(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H^T(T^{-1/2}N^{a/2}\boldsymbol{B}^T\boldsymbol{\epsilon})(1+o_P(1)).
\end{aligned}
$$

Hence I_1 is similar to F_3 in (1.36) and I_3 is similar to F_4 in (1.37) in the proof of Theorem 3, except that $\boldsymbol{H}_{20}-\boldsymbol{H}_{10}$ is now restricted to those columns with indices in H only. Using exactly the same lines of proof as in Theorem 3, we can conclude that

$$T^{1/2}\boldsymbol{\Sigma}_3^{-1/2}(I_1+I_3)\xrightarrow{\mathcal{D}}\boldsymbol{N}(\boldsymbol{0},\boldsymbol{I}_{|H|}), \tag{1.42}$$

where $\boldsymbol{\Sigma}_3=\boldsymbol{M}_3(\mathbf{S}_1+\mathbf{S}_2-\mathbf{S}_3-\mathbf{S}_3^T)\boldsymbol{M}_3^T$, with $\boldsymbol{M}_3=[(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H^T(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H]^{-1}(\boldsymbol{H}_{20}-\boldsymbol{H}_{10})_H^T$. By Assumptions R4 and R5, we can show that then I_1+I_3 is exactly $T^{1/2}N^{(1+a-b)/2}$-convergent. Since $0<a,b<1$, it is not difficult to see that I_2 is dominated by I_1+I_3 then. Also, Assumption R7 ensures $\big\|I_0\big\|_{\max}=o_P(\big\|\tilde{\boldsymbol{\delta}}-\boldsymbol{\delta}\big\|_{\max})$. All these imply that

$$T^{1/2}\boldsymbol{\Sigma}_3^{-1/2}(\tilde{\boldsymbol{\delta}}_H-\boldsymbol{\delta}_H)\xrightarrow{\mathcal{D}}\boldsymbol{N}(\boldsymbol{0},\boldsymbol{I}_{|H|}),$$

which is the asymptotic normality result we need, if we can also show that the second inequality in (1.41) is true.

From the above, since $\big\|I_0\big\|_{\max}$, $\big\|I_1\big\|_{\max}$, $\big\|I_2\big\|_{\max}$ and $\big\|I_3\big\|_{\max}$ are all $o_P(1)$, we have $\text{sign}(\tilde{\boldsymbol{\delta}}_H)=\text{sign}(\boldsymbol{\delta}_H)$. It remains to show the second inequality in (1.41).

To this end, we can show from previous results that

$$
\begin{aligned}
&\|T^{-1}(\boldsymbol{H}_{H_c}-\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_{H_c})^T\boldsymbol{B}^T\boldsymbol{X}_{\boldsymbol{\beta}-\boldsymbol{\beta}(\boldsymbol{\delta})}\text{vec}(\boldsymbol{I}_N)+T^{-1}(\boldsymbol{H}_{H_c}-\boldsymbol{B}^T\boldsymbol{Z}\boldsymbol{V}_{H_c})^T\boldsymbol{B}^T\boldsymbol{\epsilon}\|_{\max} \\
&=O_p(T^{-1/2}N^{(1+b-a)/2}),
\end{aligned}
$$

while the right hand side of the second inequality has a minimum value of

$$\frac{\gamma_T}{\|\tilde{\boldsymbol{\delta}}_{H_c}\|_{\max}}\geq\frac{\gamma_T}{\|\tilde{\boldsymbol{\delta}}_{H_c}-\boldsymbol{\delta}_{H_c}\|_{\max}}.$$

Hence, it is sufficient to prove

$$(T^{-1/2}N^{(1+b-a)/2})(\|\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_{\max}) = o_p(\lambda_T).$$

But the left hand side above has rate $T^{-1}N^{b-a} = o(\lambda_T)$ by Assumption R7. This completes the proof of the theorem. $\square$

Chapter 2

Inference for Spatial Dynamic Panel Model with different Spatial Dependence Characterizations

2.1 Introduction

Cross-sectional correlation attracts considerable attentions in both modelling and statistical inference. Spatial autoregressive model first proposed in Cliff and Ord (1973) has been used widely in many applications. The spatial effect in spatial model is measured by spatial weight matrix whose (i,j)th element reflects the strength of cross-sectional relationship between unit i and j. Regularly, the inverse of geographic or economic distance is used as the measurement and sometimes the spatial weight matrix element can also be binary, being 1 if two units are contiguous and 0 otherwise.

It is clear that spatial weight matrix is the key in spatial autoregressive model and is usually assumed as a known prior. Applied researchers often base on their subject knowledge in pre-setting such a matrix, which can be inaccurate and arbitrary at times. Recognizing the importance of correct specification of a spatial weight matrix in the accuracy of estimated model parameters, a surge of papers have focused on estimating a spatial weight matrix using data. These include the use of Lasso or the adaptive Lasso to perform sparse estimation of the spatial weight matrix in a spatial lag model like Ahrens and Bhattacharjee (2015b), Lam and Souza (2014, 2015b), the use of models with several spatial weight matrices like

Arnold et al. (2013b), Badinger and Egger (2011), Lee and Liu (2010a), LeSage and Pace (2008), or both Lam and Souza (2016b). Koroglu and Sun (2016b) proposes a nonparametric generalized method of moments to estimate a smooth function of known distance-type variables encompassing an unknown spatial weight matrix. Yet there can be more than one type of spatial weight matrices, and some may not be characterized by smooth functions of distance-type variables (e.g., negative spatial autocorrelations among some, but not all close neighbours, like policy competitions), and interpretability of spatial effects using an estimated nonparametric function can be reduced.

As some works briefly introduced above, an important tool applied recently to avoid the misspecification of spatial weight matrix in spatial model and to remain the model efficiency at the same time is the inclusion of high order spatial model where the spatial weight matrix is a linear combination of pre-specified spatial weight matrices. The multiple spatial weight matrices can capture contiguity of units in various dimensions. For instance, Tao (2005) introduces a strategic interaction model to local school expenditure using two specified spatial weight matrices based on the geographical contiguity and the economic similarity. Another perspective for using high order models is stated in Ullah (1998) that it is alternatives of a poorly specified weight matrix rather than a realistic data generating process. Therefore, a combination of pre-specified spatial weight matrices makes the model flexible to capture the complicated temporal correlations in the real data and provides the chance to identify interesting economic phenomena from the data, which is also shown in our real data analysis in Section 2.7.4.

Based on the high order spatial model above, we consider a spatial dynamic panel model allowing for both contemporaneous and time-lagged spatial correlations. The two types of spatial correlations are estimated through two different linear combinations of pre-specified spatial weight matrices. The similar study has seen a lot of theoretical advancements in recent years. The instrumental variables estimation in Gupta and Robinson (2015), Lam and Qian (2017), the generalized method of moments in Kapoor et al. (2007), Lee and Yu (2014b) and the quasi-maximum likelihood estimation in Lee and Yu (2010), Yu et al. (2008) are three approaches that are extensively developed. Obviously, instrumental variable estimation has computational simplicity but works only with the existence of instrumental variables. As for the general method of moments, as stated in Li (2017), the number of the moment conditions increases dramatically when T is large or moderately large, making itself suffer from the so-called "many moments bias".

We extend the quasi-maximum likelihood estimation technique in estimating a spatial dynamic panel data model which allows for different spatial weight matrices

representing the contemporaneous and time-lagged spatial dependence. Individual fixed effects are absorbed into our covariates (see Section 2.2). Some similar works for quasi-maximum likelihood estimation are Lee (2004), Yu et al. (2008) and Li (2017), to name but a few. To improve the computational feasibility of quasi-maximum likelihood estimation, the instrumental-like variable estimation proposed in Lam and Qian (2017) is applied as the initial value in the optimization of quasi-maximum likelihood estimation. In Lam and Qian (2017), the exogenous variables are needed to be instrumental variables. We use the regressors in the proposed model as instrumental variables even they are not perfectly exogenous. It is shown in Section 2.7 that this initial value performs well in practice.

Our main contributions are not only the extension of quasi-maximum likelihood estimation but also the inference for spatial autoregressive model, which is rarely discussed in literature. Firstly, we provide a unified spatial test statistic for testing the presence of either contemporaneous or time-lagged spatial correlation, or both. Although asymptotic normality results are given for quasi-maximum likelihood estimation in Yu et al. (2008), for general method of moments estimation in Lee and Liu (2010a) and for instrumental variable estimation in Gupta and Robinson (2015), the following tests based on the asymptotic normality results are not mentioned in the above researches. Baltagi et al. (2007) derives several Lagrange multiplier tests for the panel data regresssion model only containing contemporaneous spatial effect on disturbance, while our work also contains the dynamic spatial effects. More importantly, as the proposed model is high order spatial model, which applies a linear combination of specified spatial weight matrices for spatial effect, this unified test can be used to test whether all of them are necessary.

Another important inference is a diagnostic test for testing if the fitted residuals are white noise. Chang et al. (2017) proposes this test by approximating the distribution of the maximum absolute auto-/cross-correlations of the component series. This diagnostic test allows for the dimension N of the fitted residual vector to be growing with the sample size T, same as the setting of our spatial model. To the best of our knowledge, in spatial econometrics context, it is the first time that this high dimensional diagnostic test is applied for model fitness checking, which is very fundamental in real data analysis. Therefore, our test does not only test the whiteness of the fitted residuals themselves, but can also improve the trust on the estimated contemporaneous and time-lagged spatial effect.

The rest of this chapter is organized as follows. Section 2.2 presents our model and a sufficient condition for stationarity. Section 2.3 describes several areas where our model can be useful and how it is a generalization of commonly used models. The quasi-maximum likelihood estimation, together with necessary assumptions

and asymptotic results for our estimators are presented in Section 2.4. Section 2.5 presents the Wald statistic and the corresponding asymptotic theory for testing different spatial correlations, while Section 2.6 provides the test statistic and presents the asymptotic theory for the high dimensional diagnostic test for white noise in our model. Section 2.7 presents our simulation results as well as a stock return analysis. All technical proofs are relegated to Section 2.10, with Appendix A detailing how we find the derivatives our the log-likelihood function of our model.

2.2 The model

For $t = 1, \ldots, T$, we consider the model

$$y_t = \big(\sum_{i=1}^{M} \alpha_{0i}\mathbf{W}_{0i}\big)y_t + \big(\sum_{i=1}^{M} \gamma_{0i}\mathbf{W}_{0i}\big)y_{t-1} + \phi_0 y_{t-1} + \mathbf{X}_t\beta_0 + \boldsymbol{\epsilon}_t, \tag{2.1}$$

where $y_t = (y_{1t}, \ldots, y_{Nt})^T$ and $\boldsymbol{\epsilon}_t = (e_{1t}, \ldots, e_{Nt})^T$ are N dimensional vectors, e_{it} is an innovation series with zero mean and variance σ_0^2, $\{\mathbf{W}_{0i}, i = 1, \ldots, M\}$ are $N \times N$ predetermined spatial weight matrices. Finally, $\mathbf{X}_t$ is an $N \times \kappa_x$ matrix of non-stochastic regressors. The proposed model here is a special case of model (1.1) in Chapter 1 when we set $p = 1$. As the main purpose of this Chapter is for inference in spatial modelling, we set $p = 1$ for simplicity.

Same as discussed in Section (1.2.1), the zero diagonal elements in all $\mathbf{W}_{0i}$ make the dynamic and contemporaneous effect between different units in panel capture by α_{0i} and γ_{0i} and the dynamic interaction for same unit is shown in ϕ_0.

Model (2.1) is different from a traditional spatial dynamic panel model in the following two aspects. First, we consider a general spatial dependence, which can be decomposed as a linear combination of pre-specified "spatial dependence", called spatial weight matrices. We then estimate this linear combination through data, so that information from traditional expert knowledge on spatial relationship can be merged with microstructure of data to enhance the accuracy of the resulting spatial dependence structure.

In Li (2017), a more specific explanation of the need for high order spatial models is given by a showcase example. If there are M groups in N spatial units and the spillover effects only exist within groups, a small matrix $\tilde{\mathbf{W}}_m$, $m = 1, \ldots, M$ is denoted as the spatial interaction within group m. In classical model, a block diagonal matrix $\mathbf{W} = \text{diag}(\tilde{\mathbf{W}}_1, \ldots, \tilde{\mathbf{W}}_M)$ is applied, which makes the spillover effects same across groups. However, if we apply a high order spatial model using a

linear combination of $\mathbf{W}_m = \text{diag}(0, \ldots, \tilde{\mathbf{W}}_m, \ldots, 0)$ to fit the data, we can capture the group-dependent strengths of spatial spillover effects. The examples for this model can also be found in Lacombe (2004), McMillen et al. (2007) among others.

Secondly, to reduce the parameter dimension, we model the fixed effects μ_{0i} by $\mu_{0i} = \mu_0 + Z_i'\eta$, where μ_0 is the intercept shared by all the components and Z_i is a $p-$dimensional component specific covariate vector, where p is a fixed integer. We then embed this into the covariate matrix $\mathbf{X}_t$, so that model (2.1) does include the fixed effects for all the components in y_t. As for this elimination of individual effect, Section 2.8 introduces some other ideas to avoid the incidental parameter problem, which can also easily to be applied into our proposed model.

Denote $\alpha = (\alpha_1, \cdots, \alpha_M)' \in \Lambda$, with true value $\alpha_0 = (\alpha_{10}, \cdots, \alpha_{M0})'$, and $\gamma = (\gamma_1, \cdots, \gamma_M)' \in \Gamma$, with true value $\gamma_0 = (\gamma_{10}, \cdots, \gamma_{M0})'$. For convenience of our presentation, we also define $\varphi = (\gamma', \phi)'$, $\delta = (\gamma', \phi, \beta')'$, $\xi = (\alpha', \delta')'$ and $\theta = (\xi', \sigma^2)'$, with similar definitions for their respective true values. θ is the parameters to be estimated in this chapter. Define

$$H_N(\alpha) = I_N - \sum_{i=1}^{M} \alpha_i \mathbf{W}_{0i}, \quad \Upsilon_N(\varphi) = \sum_{i=1}^{M} \gamma_i \mathbf{W}_{0i} + \phi I_N, \quad A_N(\alpha, \varphi) = H_N^{-1}(\alpha)\Upsilon_N(\varphi).$$

Then, y_t can be rewritten as

$$y_t = A_N y_{t-1} + H_N^{-1}\mathbf{X}_t\beta_0 + H_N^{-1}\boldsymbol{\epsilon}_t, \tag{2.2}$$

where $H_N = H_N(\alpha_0)$ and $A_N = A_N(\alpha_0, \varphi_0)$.

From (2.2), model (2.1) can be embedded into the framework of a vector autoregressive (VAR) model. By Corollary (5.6.16) of Horn and Johnson (2012), a sufficient condition for the stationarity of model (2.1) is $\|A_N\| < 1$, where the matrix norm can be either the L_1 or L_∞ norm, defined by $\|M\|_1 = \max_j \sum_i |m_{ij}|$ and $\|M\|_\infty = \max_i \sum_j |m_{ij}|$ respectively. If all $\mathbf{W}_{0i}$ are row-standardized, in the sense that the sum of the absolute value of the elements in each row of $\mathbf{W}_{0i}$ equals to 1 (see for example, LeSage and Pace (2009) for the use of such row-standardization), then we have $\|\sum_{i=1}^{M} \alpha_{0i}\mathbf{W}_{0i}\|_\infty \leq \sum_{i=1}^{M} \alpha_{0i}$. By Lemma 2.3.3 of Golub and van Loan (1996), we obtain that

$$\|H_N^{-1}\|_\infty \leq \frac{1}{1 - \sum_{i=1}^{M} \alpha_{0i}}.$$

Note that

$$\|A_N\|_\infty \leq \|H_N^{-1}\|_\infty \|\Upsilon_N\|_\infty \leq \frac{\sum_{i=1}^{M} |\gamma_{0i}| + |\phi_0|}{1 - \sum_{i=1}^{M} |\alpha_{0i}|}.$$

Therefore, $\sum_{i=1}^{M} |\alpha_{0i}| + \sum_{i=1}^{M} |\gamma_{0i}| + |\phi_0| < 1$ is a sufficient condition for $\|A_N\|_\infty < 1$, and hence stationarity of model (2.2), which is assumed throughout the chapter.

2.3 Some application examples

As known, the dynamic spatial model is important in many fields. Therefore, the proposed model can be very useful in a variety of economic and social setups. In the following, we give three examples.

1. ***Economics***. Economic activities are often concentrated geographically. The development of policies for strengthening economic growth, for example, is therefore dependent on the understanding of the clustering of such activities spatially. Spatial econometric models, through the specification of a spatial weight matrix, is designed to account for spatial interactions among observed units, allowing for the effects of exogenous variables.

   An example in studying the growth of different European regions is Baumont et al. (2003). In it, different distance-based spatial weight matrices W are explored, and plots of log per-capital GDP for different European countries (Y) against the spatial-lagged ones (WY) consistently reveal two clusters of regions in Europe, at the same time showing a consistent positive slope, called the spatial correlation. In essence, it suggests the use of a model of the form

   $$Y = \rho WY + X\beta + \epsilon,$$

   where ρ is the spatial correlation, and X is a matrix of exogenous covariates, can be useful. Such a spatial lag model is discussed in details in Anselin (2010).

   While the model can be useful, it is generally agreed that the specification of W has to be given extreme care, since different W can give substantially different results. Abreu et al. (2005) suggests that no matter a contiguity matrix (where neighboring regions are coded 1 and others as 0) or distance-based matrices are used, they should arise from underlying theoretical considerations, on top of being exogenous from the variables that are being studied. This creates

difficulties in practice since even for distance based matrices, using d^{-1} or d^{-2} can be potentially different. Indeed, Baumont et al. (2003) explores different distance-based spatial weight matrices before coming to their conclusion of regional polarization.

The model (2.1) we proposed can handle this situation nicely by considering a linear combination of some sensible spatial weight matrices, and let the "best" spatial weight matrix to be estimated from the data as the "best" linear combination. Lee and Liu (2010a) also considers a model with a linear combination of spatial weight matrices, and call this higher order spatial lags, with generalized method of moments proposed to estimate the model. Our model can be considered a generalization of theirs since we also include a time and spatial-lagged term $(\phi_0+\sum_{i=1}^{M}\gamma_{0i}\mathbf{W}_{0i})y_{t-1}$. In the case of regional growth by country for instance, a time-lagged shock from neighboring regions of a country can still carry an effect to the current GDP growth of the particular country. Such a shock can be well estimated by our model with such a time and spatial-lagged term.

2. ***Social Network***. Social network represents general relationships among individual units. A prominent example is the Twitter or Facebook network, where an individual is linked to another person if he or she is "following" or "friend" of another. In Bramoullé et al. (2009), the level of a student's recreational activity y_i is modeled as

$$y_i = \alpha + \beta\frac{\sum_{j\in P_i} y_j}{N_i} + \gamma x_i + \delta\frac{\sum_{j\in P_i} x_j}{N_i} + \epsilon_i,\ E(\epsilon_i|\mathbf{x}) = 0,$$

where P_i is the set of friends for individual i with $N_i = |P_i|$, and x_j is the parents' income for the jth individual. The model clearly includes the mean of the friends' level as an endogenous factor for the recreational level of individual i, which can be considered a peer effect. His or her parents' income, as well as the mean parents' income of all friends, are also intuitively important factors that are included in the model above. This particular social network model can actually be captured by our model with $T = r = 1$, $\alpha_{01} = \beta$, $\gamma_{01} = \phi_0 = 0$, $W_{N1} = (1/N_i \cdot 1_{\{j\in P_i\}})$ and

$$X_{N1} = \begin{pmatrix} 1 & x_1 & x_1/N_1 \cdot 1_{\{1\in P_1\}} & \cdots & x_N/N_1 \cdot 1_{\{N\in P_1\}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_1/N_N \cdot 1_{\{1\in P_N\}} & \cdots & x_N/N_N \cdot 1_{\{N\in P_N\}} \end{pmatrix},\ \beta_0 = (\alpha, \gamma, \delta\mathbf{1}_N')',$$

where $\mathbf{1}_N$ is a column vector of N ones, and 1_A is 1 under the event A and 0 otherwise. If $N_i < \infty$, each non-zero entry in the ith row of X_{N1} (apart from the constant 1) can be the entries in the covariate vector Z_i in our model (with some entries possibly 0 if the N_i's are not all equal), with $p = \max(N_i) + 1$ and $\eta = (\gamma, \delta \mathbf{1}'_{\max(N_i)})'$. Clearly, our model is more general than that introduced in Bramoullé et al. (2009), with the possibility of adding time-lagged peer effects.

Our model is also a generalization of the network vector autoregression proposed by Zhu et al. (2017). Compared to the model (2.1) in their paper, we have the spatial interaction term $\sum_{i=1}^{M} \alpha_{0i} \mathbf{W}_{0i}$, which can be important to capture the endogenous peer effects in a social network. At the same time, setting $W_{N1} = (a_{ij}/N_i)$, $\gamma_{01} = \beta_1$ and $\phi_0 = \beta_2$ in our model, we have the autoregressive term in model (2.1) of Zhu et al. (2017). The analysis of our paper is also significantly different from that in Zhu et al. (2017) because of the endogenous peer effects term.

3. ***Finance***. Spatial dependence can be used in different areas of finance for interpretation/forecasting purposes. One important interest lies in the spatial dependence among different financial markets. This is closely related to the contagion of financial markets, and hence how world market shocks are propagated to one another. In Fernandez (2011), a spatial Capital Asset Pricing Model is used for modeling asset returns in several emerging markets, where such a model is in parallel to the spatial lag model, giving an extra risk term coming from a weighted average of neighboring firms on top of the usual risk premium from the market portfolio. In doing so, four different spatial weight matrices are calculated based on four different financial indicators. They are market capitalization (relative to firm size), the market-to-book, the dividend yield and the debt maturity ratios. Subsequent analyses use these spatial weight matrices separately for drawing their conclusions. Our model, however, can accommodate these different spatial weight matrix specifications by finding a linear combination of them that is more adapted to the data. This can also enhance any conclusions that can be drawn only separately from using different spatial weight matrices individually.

   Indeed, Arnold et al. (2013b) has used a linear combination of spatial weight matrices when using a spatial lag model without exogenous covariates to model stock returns. In their notations, they model the vector of stock returns y_t by

$$y_t = \rho_g W_g y_t + \rho_b W_b y_t + \rho_l W_l y_t + \epsilon_t,$$

where W_g, W_b and W_l are spatial weight matrices such that $W_g y_t$ represents the weighted market return of the same day, $W_b y_t$ the weighted market return of the respective industrial branches, and finally $W_l y_t$ the weighted local market return of the respective countries. This simple model already gives very good Value-at-Risk forecasts. The model we proposed certainly generalizes theirs with the possibility of adding important exogenous covariates, as well as time-lagged spatial dependence.

2.4 The quasi-maximum likelihood estimators

Different from the Chapter 1, quasi-maximum likelihood estimators are applied in this Chapter. As discussed in Introduction, instrumental variable used in profiled least square estimation is hard to find in real data analysis and the model (2.1) sets $p = 1$ to improve the computational feasibility of the quasi-maximum likelihood estimators.

The total number of parameters of model (2.1) is $2M + \kappa_x + 2$ including α, γ, ϕ, β and σ. The log-likelihood function of model (2.1) is

$$\log L_{NT}(\theta) = -\frac{NT}{2}\log 2\pi - \frac{NT}{2}\log\sigma^2 + T\log|H_N(\alpha)| - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\boldsymbol{\epsilon}_t'(\xi)\boldsymbol{\epsilon}_t(\xi), \quad (2.3)$$

where

$$\boldsymbol{\epsilon}_t(\xi) = H_N(\alpha)y_t - \Upsilon_N(\varphi)y_{t-1} - \mathbf{X}_t\beta. \quad (2.4)$$

The quasi-maximum likelihood estimator (QMLE) $\hat{\theta}_{NT}$ is defined as

$$\hat{\theta}_{NT} = \arg\min_{\theta\in\Theta} l_{NT}(\theta),$$

where Θ is the parameter space and

$$l_{NT}(\theta) = -\frac{1}{NT}\log L_{NT}(\theta). \quad (2.5)$$

To obtain the asymptotic results of QMLE $\hat{\theta}_{NT}$, we need the following assumptions. First, we introduce some notations. Let

$$L_{Nt} = (\mathbf{W}_{01}y_t, \mathbf{W}_{02}y_t, \cdots, \mathbf{W}_{0M}y_t) \quad \text{and} \quad O_{Nt} = (L_{N,t-1}, y_{t-1}, \mathbf{X}_t) \quad (2.6)$$

and we denote $B_{Ni} = \mathbf{W}_{0i} H_N^{-1}$ and define

$$\tilde{O}_{Nt} = (B_{N1} O_{Nt} \delta_0, \cdots, B_{NM} O_{Nt} \delta_0) \text{ and } \tilde{tr}(B_N) = (tr(B_{N1}), \cdots, tr(B_{NM}))'. \quad (2.7)$$

A1. The spatial weight matrix $\mathbf{W}_{0i}$ is a constant, with zero diagonal elements for each $i = 1, \ldots, M$.

A2. The error series $\{e_{it}\}$ are independent across i and are martingale differences across t, with $Ee_{it} = Ee_{it}^3 = 0$ and $Ee_{it}^2 = \sigma_0^2$. Furthermore,

(i) there exists a constant $\tau > 0$ such that $E|e_{it}|^{4+\tau} < \infty$;

(ii) the quantity $\sum_{h=0}^{\infty} \Gamma(h)$ is finite, with $\Gamma(h) = \text{cov}(e_{it}^2, e_{i,t+h}^2)$.

A3. The matrix $H_N(\alpha)$ is invertible for all $\alpha \in \Lambda$, where Λ is compact and α_0 is in the interior of Λ.

A4. The elements of $\mathbf{X}_t$ are non-stochastic and bounded uniformly in N and T. Furthermore, $\lim_{T\to\infty} \frac{1}{NT} \sum_{t=1}^{T} \mathbf{X}_t' \mathbf{X}_t$ exists and is nonsingular.

A5. The quantities $\|\mathbf{W}_{0i}\|_1$ and $\|\mathbf{W}_{0i}\|_\infty$ are both bounded uniformly in N for $i = 1, \ldots, M$. The same goes for $\|H_N^{-1}(\alpha)\|_1$, $\|H_N^{-1}(\alpha)\|_\infty$.

A6. The quantities $\|\sum_{h=1}^{\infty} \text{abs}(A_N^H)\|_1$ or $\|\sum_{h=1}^{\infty} \text{abs}(A_N^H)\|_\infty$ are bounded, where $\text{abs}(A)$ is the matrix of absolute values of the entries in A.

A7. The dimension N is a nondecreasing function of T as T goes to infinity.

A8. The matrix $\lim_{T\to\infty} E\mathcal{J}_{NT}$ is nonsingular, where $\mathcal{J}_{NT} = \frac{1}{NT} \sum_{t=1}^{T} (\tilde{O}_{Nt}, O_{Nt})'(\tilde{O}_{Nt}, O_{Nt})$.

Remark 1 Assumption A1 is standard for spatial econometric models. Assumption A2 allows for dependence of the e_{it}'s at different time points, which relaxes the usual independence and identical distribution assumptions for consistency. The requirement $Ee_{it}^3 = 0$ is a technical condition for a more explicit expression for the global identification of the parameters. Assumption A3 and A6 ensure that model (2.1) has a moving average representation, and compactness is a technical condition. Assumption A4 is for the convenience of proof and can be relaxed to $\mathbf{X}_t$ being stochastic with some moments restrictions. Assumption A5 limits the spatial correlation to a manageable degree. Assumption A6 limits the dependence of y_t across time and space, and together with Assumption A2, A3 and A5 imply that model (2.1) has the moving average representation

$$y_t = \sum_{h=0}^{\infty} A_N^h H_N^{-1} (\boldsymbol{\epsilon}_{t-h} + \mathbf{X}_{t-h} \beta_0) = \mathcal{E}_{Nt} + \mathcal{X}_{Nt}, \quad (2.8)$$

where $\mathcal{E}_{Nt} = \sum_{h=0}^{\infty} A_N^h H_N^{-1} \boldsymbol{\epsilon}_{t-h}$ and $\mathcal{X}_{Nt} = \sum_{h=0}^{\infty} A_N^h H_N^{-1} \mathbf{X}_{t-h}$. Assumption A7 includes the fixed N and $N \to \infty$ scenarios. Finally, Assumption A8 is an identification condition for the parameters of model (2.1).

As the first step for the inference in our spatial model, we straightly give the following Theorems 1 to 3 based on Yu et al. (2008). As we discussed in Introduction, it is interesting to use these inference ideas to test the spatial modelling fitness. As for Theorem 4, we can have a diagnostic testing for the fitted residuals in our model being white noise.

Theorem 1. *Suppose Assumption A1-A8 hold, then θ_0 is globally identified and $\hat{\theta}_{NT} \xrightarrow{P} \theta_0$ with increasing N and T.*

Theorem 2. *Suppose the conditions of Theorem 1 hold. If furthermore e_{it}^2 are uncorrelated across t, we have*

$$\sqrt{NT}(\hat{\theta}_{NT} - \theta_0) \xrightarrow{d} N(0, \Sigma^{-1}\Omega\Sigma^{-1}),$$

where

$$\Omega = \lim_{T,N\to\infty} NT \cdot E\Big(\frac{\partial l_{NT}(\theta_0)}{\partial \theta} \frac{\partial l_{NT}(\theta_0)}{\partial \theta'}\Big), \quad \Sigma = \lim_{T,N\to\infty} E\Big(\frac{\partial^2 L_{NT}(\theta_0)}{\partial \theta \partial \theta'}\Big).$$

Theorem 2 means that the QMLE is asymptotic normal with standard rate $\sqrt{NT}$ even if N is much larger than T. This is because we model the fix effects by some specific variables to reduce the dimension of the parameter to a fixed number.

2.5 The tests for spatial autocorrelation

It is very important to access whether each specification of the spatial weight matrices used in the model is appropriate for the data. Also, testing whether the data has contemporaneous or time-lagged spatial interaction is significant in spatial modelling. This section considers a unified framework for testing the linear relationship of the coefficients:

$$H_0 : a'\theta_0 = c \text{ against } H_1 : a'\theta_0 \neq c, \tag{2.9}$$

where a is a known $(2M + 2 + \kappa_x) \times d$-dimensional constant matrix and c is a d-dimensional known constant vector. With different values of a, (2.9) includes

different tests for spatial effects in the data. We can list the following hypothesis as examples,

(i) $H_0^1 : \alpha_{0i} = 0,\ i = 1, \ldots, M$. Under the null hypothesis, there is no contemporaneous spatial correlation.

(ii) $H_0^2 : \gamma_{0i} = 0,\ i = 1, \ldots, M$. Under the null hypothesis, there is no lag-1 spatial correlation.

(iii) $H_0^3 : \alpha_{0i} = 0,\ \gamma_{0i} = 0,\ i = 1, \ldots, M$. Under the null hypothesis, there is no spatial correlation at all.

Denote

$$\hat{\Sigma}_{NT} = \frac{\partial^2 l_{NT}(\hat{\theta}_{NT})}{\partial\theta\partial\theta'}, \quad \hat{\Omega}_{NT} = NT\frac{\partial l_{NT}(\hat{\theta}_{NT})}{\partial\theta}\frac{\partial l_{NT}(\hat{\theta}_{NT})}{\partial\theta}.$$

We use a Wald test statistic for testing H_0 against H_1 in (2.9), which is defined as

$$\mathcal{W} = NT(a'\hat{\theta}_{NT} - c)'\big(a\hat{\Sigma}_{NT}^{-1}\hat{\Omega}_{NT}\hat{\Sigma}_{NT}^{-1}a'\big)^{-1}(a'\hat{\theta}_{NT} - c).$$

Based on the asymptotic results in the last section, we have the following theorem.

Theorem 3. *Suppose the conditions of Theorem 2 hold. Then* $\mathcal{W} \xrightarrow{d} \chi^2(d)$ *under* H_0 *as* $T, N \to \infty$.

2.6 Diagnostic testing for the model

Another more important test is a diagnostic testing for the fitted residuals in model (2.1) being white noise, as

$$H_0 : \{\boldsymbol{\epsilon}_t\} \text{ is white noise} \quad \text{v.s.} \quad H_1 : \{\boldsymbol{\epsilon}_t\} \text{ is not white noise.}$$

Ignoring the testing on the residuals results in consistent but inefficient estimates of the regression coefficients and biased standards errors, see Baltagi (2008). Baltagi et al. (2007) adds the correlation in the error structure, which is similar to spatial disturbance autoregressive model, and derives the tests based on this specified model. We adopt the test based on the maximum cross-correlations proposed in Chang et al. (2017) to test whether the fitted residuals are white noise vector.

We introduce some notations first before giving the test statistic. Denote the autocovariance and autocorrelation matrix of $\boldsymbol{\epsilon}_t$ at lag k by $\Lambda(k)$ and $\rho(k)$ respectively, where

$$\Lambda(k) = \mathrm{Cov}(\boldsymbol{\epsilon}_{t+k}, \boldsymbol{\epsilon}_t), \quad \rho(k) = \mathrm{diag}(\Lambda(0))^{-1/2}\Lambda(k)\mathrm{diag}(\Lambda(0))^{-1/2},$$

with $\mathrm{diag}(A)$ representing the diagonal matrix with diagonal elements of A. Let

$$\hat{\rho}(k) = (\hat{\rho}_{ij}(k)) = \mathrm{diag}(\hat{\Lambda}(0))^{-1/2}\hat{\Lambda}(k)\mathrm{diag}(\hat{\Lambda}(0))^{-1/2}$$

be the sample lag-k autocorrelation matrix for the estimated errors $\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_t(\hat{\xi}_{NT})$ (recall that $\hat{\xi}_{NT}$ is a component vector of $\hat{\theta}_{NT}$ defined in Section 2.4), where

$$\hat{\Lambda}(k) = \frac{1}{T}\sum_{t=1}^{T-k} \hat{\boldsymbol{\epsilon}}_{t+k}\hat{\boldsymbol{\epsilon}}_t'. \tag{2.10}$$

The test statistic is then defined as

$$\mathcal{T}_{NT} = \max_{1\le k\le K} \mathcal{T}_{N,k},$$

where $\mathcal{T}_{N,k} = \max_{1\le i,j\le N}\sqrt{T}|\hat{\rho}_{ij}(k)|$ and K is a prescribed positive integer. We reject H_0 if $\mathcal{T}_{NT} > c_\eta$, where c_η is determined by

$$Pr(\mathcal{T}_{NT} > c_\eta | H_0) = \eta.$$

Unfortunately, we do not know the asymptotic distribution of $\mathcal{T}_{NT}$, which yields the critical value c_η. Following the lines of Chang et al. (2017), we use the distribution of $\|\mathcal{G}\|_\infty$ to approximate the distribution of $\mathcal{T}_{NT}$, where $\mathcal{G}$ is a multivariate normal random vector with mean 0 and covariance matrix whose estimator is $\hat{\Psi}_{NT}$. The matrix $\hat{\Psi}_{NT}$ is defined by

$$\hat{\Psi}_{NT} = (I_K \otimes \hat{\mathcal{Z}})\mathcal{Q}_{NT}(I_K \otimes \hat{\mathcal{Z}}),$$

where $\hat{\mathcal{Z}} = \left[\mathrm{diag}(\hat{\Lambda}(0))\right]^{-1/2} \otimes \left[\mathrm{diag}(\hat{\Lambda}(0))\right]^{-1/2}$, and $\mathcal{Q}_{NT}$ is defined by

$$\mathcal{Q}_{NT} = \sum_{j=-T+K+1}^{T-K-1} \mathcal{K}(\frac{j}{b_T})\hat{\mathcal{H}}(j),$$

where

$$f_t = \big(\mathrm{vec}(\hat{\boldsymbol{\epsilon}}_{t+1}\hat{\boldsymbol{\epsilon}}_t'), \cdots, \mathrm{vec}(\hat{\boldsymbol{\epsilon}}_{t+K}\hat{\boldsymbol{\epsilon}}_t')\big)',$$

$$\hat{\mathcal{H}}(j) = \begin{cases} \frac{1}{T-K}\sum_{t=j}^{T-K} f_t f_{t-j}', & j \geq 0 \\ \frac{1}{T-K}\sum_{t=-j+1}^{T-K} f_{t+j} f_t', & j < 0, \end{cases}$$

$$\mathcal{K}(x) = \frac{25}{12\pi^2 x^2}\Big[\frac{\sin(6\pi x/5)}{6\pi x/5} - \cos(6\pi x/5)\Big],$$

with b_T a bandwidth diverging with T (see Section 2.7.2 for more details about the data-driven bandwidth). Then we define $\hat{c}_\eta$ by

$$Pr(\|\hat{\mathcal{G}}\|_\infty > \hat{c}_\eta | \hat{\boldsymbol{\epsilon}}_1, \ldots, \hat{\boldsymbol{\epsilon}}_T) = \eta,$$

where $\hat{\mathcal{G}} \sim N(0, \hat{\Psi}_{NT})$. The value $\hat{c}_\eta$ can serve as the critical value for $\mathcal{T}_{NT}$ due to the following theorem, which needs the additional assumptions as below,

A9. There exist constants c_1, $c_2 > 0$ and $\tilde{r}_1 \in (0, 2]$ such that $Pr(|e_{11}| > x) \leq c_1 \exp(-c_2 x^{\tilde{r}_1})$.

A10. The error $\boldsymbol{\epsilon}_t$ in model (2.1) is β-mixing with mixing coefficients satisfying $\tilde{\beta}_k \leq \exp(-c_3 k^{\tilde{r}_2})$ for some constants $c_3 > 0$ and $\tilde{r}_2 \in (0, 1]$.

A11. There exists a constant $c_4 > 0$ and $\zeta > 0$ such that

$$\begin{aligned} &\tfrac{1}{c_4} \liminf_{q\to\infty} \inf_{m\leq 0} E\big(|\tfrac{1}{\sqrt{q}}\textstyle\sum_{t=m+1}^{m+q} e_{i,t+k}e_{jt}|^{2+\zeta}\big) \\ \leq &\limsup_{q\to\infty} \sup_{m\leq 0} E\big(|\tfrac{1}{\sqrt{q}}\textstyle\sum_{t=m+1}^{m+q} e_{i,t+k}e_{jt}|^{2+\zeta}\big) \leq c_4. \end{aligned}$$

Theorem 4. *Suppose Assumption A9-A11 and the conditions of Theorem 2 hold. If $B_N \sim T^\rho$ for $0 < \rho < \min\{1/6, \tilde{r}_2/(1+\tilde{r}_2)\}$ and $\log(p) \leq CT^{\tilde{\rho}}$ for some constants C and $\tilde{\rho}$, then as $T \to \infty$,*

$$Pr(\mathcal{T}_{NT} > \hat{c}_\eta | H_0) \to \eta.$$

It is easy to find that the estimated errors should be a white noise vector if the model is adequate, which dose not only leave more belief on the consistency of estimated coefficients and also makes the spatial effects and the spatial weight matrices included into our model more trustable .

2.7 Numerical Study

2.7.1 Performance of QMLE

We generate $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ in model (2.1) as follows. Set $\kappa_x = 3$ (so 3 columns in $\mathbf{X}_t$) and we generate $\mathbf{X}_t$ using $\text{vec}(\mathbf{X}_t) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_X)$, where

$$\boldsymbol{\Sigma}_X = \begin{bmatrix} 2\boldsymbol{I}_N & 0.5\boldsymbol{I}_N & 0.5\boldsymbol{I}_N \\ 0.5\boldsymbol{I}_N & 2\boldsymbol{I}_N & 0.5\boldsymbol{I}_N \\ 0.5\boldsymbol{I}_N & 0.5\boldsymbol{I}_N & 2\boldsymbol{I}_N \end{bmatrix}.$$

The $\boldsymbol{\epsilon}_t$'s are independent of each other, each follows a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{I}_N)$. Unlike the corresponding simulation setting in Chapter 1, the simulation for quasi-maximum likelihood estimation needs a error covariance matrix as $\sigma^2\boldsymbol{I}_N$ to achieve the log-likelihood function 2.3. Therefore, we use $\boldsymbol{I}_N$ as covariance matrix for error term $\boldsymbol{\epsilon}_t$, which is commonly used like Lam and Souza (2018) and Yu et al. (2008).

We set $M = 3$ for model (2.1). All elements in the parameters ϕ, α, γ and β are independently generated from the uniform distribution $\boldsymbol{U}(0, 1)$ in each simulation run. To satisfy the sufficient condition for stationarity, every element in ϕ_0, α_0, γ_0 is divided by 1.2 times the absolute sum of all of them.

As $M = 3$, we need 3 specified spatial weight matrices W_{01}, W_{02} and W_{03}. To simplify the simulation procedure so that each $\mathbf{W}_{0i}$ still has eigenvalues less than 1 in magnitude, which is a part of the stationarity conditions for model (2.1), we generate each $\mathbf{W}_{0i}$ with only the first three off-diagonals (both lower and upper) being non-zero. Another procedure where we produce sparse $\mathbf{W}_{0i}$ has very similar results and is not presented in this chapter.

We use the MATLAB function `fmincon` to evaluate the quasi-maximum likelihood estimation $\hat{\theta}_{NT}$ constrained for stationarity. Since the likelihood function is in general not convex, it is necessary to find a good initial value for the procedure. We apply the method in Lam and Qian (2017), which can provide an accurate least square type estimator if suitable instrumental-type variables exist. In our simulations, the $\mathbf{X}_t$'s are exogenous and can be the "instrumental variables". We also simulate $\mathbf{X}_t$ to be correlated with $\boldsymbol{\epsilon}_t$ in general while still using the $\mathbf{X}_t$'s as the "instruments" needed in Lam and Qian (2017). It turns out that the initial values obtained are still good enough for quasi-maximum likelihood estimation $\hat{\theta}_{NT}$ to converge to reasonable values as long as the correlations between the elements in $\mathbf{X}_t$

and $\boldsymbol{\epsilon}_t$ are not too strong, and hence the simulation results are very similar in the end.

We repeat our simulation 500 times. Figure 2.1 shows the boxplots of the averaged L_1-error for $\hat{\theta}_{NT}$, which is $\|\hat{\theta}_{NT} - \theta_0\|_1/11$, under different combinations of (N, T). It is clear that $\hat{\theta}$ is convergent to the true value as N or T becomes large.

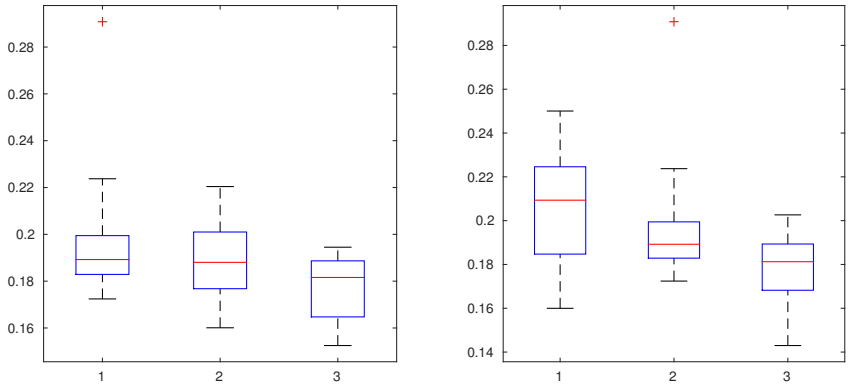


Figure 2.1: Boxplots of $\|\hat{\theta}_{NT} - \theta_0\|_1/11$. Left panel, from left to right: $N = 50, 100, 150, T = 100$. Right panel, left to right: $T = 50, 100, 150, N = 100$.

To examine the asymptotic normality of $\hat{\theta}_{NT}$ in Theorem 2, we use the previous settings for 500 simulation runs. Using the results in Theorem 2, we calculate $\widehat{\boldsymbol{\Sigma}} = E(\partial^2 l_{Nt}(\hat{\theta}_{NT})/\partial\theta\partial\theta')$ as an estimator for $\boldsymbol{\Sigma}$. We choose $\hat{\sigma}^2$ from $\hat{\theta}_{NT}$ and standardize it using the true value σ_0^2 and the corresponding entry in the estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}$ when $(N, T) = (100, 100)$. The histogram and normal probability plots in Figure 2.2 show that our standardized estimator follows a standard normal distribution. Other combinations of (N, T) produces similar results and are not shown here.

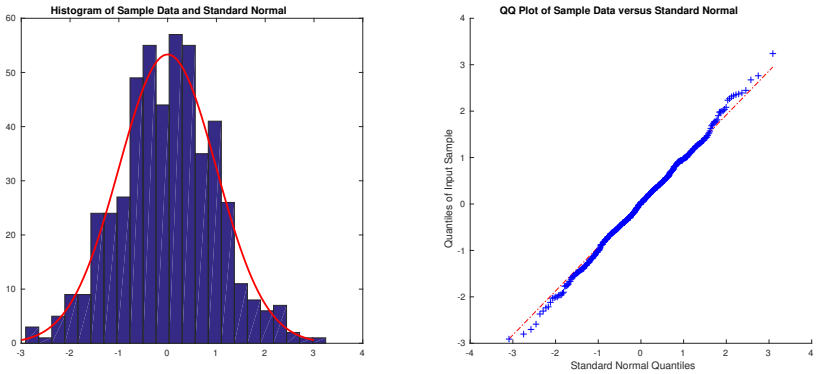


Figure 2.2: Histograms (left) and normal probability plot (right) for standardized $\hat{\sigma}^2$. Standardization used the estimated asymptotic covariance matrix derived in Theorem 2.

2.7.2 Performance of spatial and diagnostic tests

We use the settings in the previous section except for the values of α_0 and γ_0, which are set according to three different scenarios:

1. Both contemporaneous and lag-1 spatial correlations are present, $\alpha_0 \neq 0, \gamma_0 \neq 0$;

2. Only lag-1 spatial correlation presents, $\alpha_0 = 0, \gamma_0 \neq 0$;

3. Only contemporaneous spatial correlation presents, $\alpha_0 \neq 0, \gamma_0 = 0$.

Under scenario 1, all null hypotheses in Section 2.5 should be rejected, while only H_0^2 and H_0^3 should be rejected under scenario 2. Under scenario 3, only H_0^1 and H_0^3 should be rejected. Table 2.1 shows the proportion of times H_0^1, H_0^2 and H_0^3 are rejected in 500 simulation runs under the three different scenarios above. Significance level of the tests are set at 5%. It is clear that in many combinations of (N, T), our spatial test is doing very well, rejecting the hypothesis that should not be rejected below 5% of times, and at the same time rejecting those hypotheses that should be rejected usually at or above 95% of times.

For the diagnostic test in Section 2.6, since we generate our data using white noise as errors, ideally we should not reject our null hypothesis of $\hat{\boldsymbol{\epsilon}}_t$ being a white noise. To apply the diagnostic test, we use the data-driven bandwidth $b_T = 1.3221\{\hat{a}(2)T\}^{1/5}$ suggested in Section 6 of Andrews (1991) and in Section 4 of Chang et al. (2017), where $\hat{a}(2) = \{\sum_{l=1}^{N^2K} 4\hat{\rho}_l^2\hat{\phi}_l^4(1-\hat{\rho}_l)^{-8}\}\{\sum_{l=1}^{N^2K} \hat{\phi}_l^4(1-\hat{\rho}_l)^{-4}\}^{-1}$ with $\hat{\rho}_l$ and $\hat{\phi}_l^2$ being, respectively, the estimated autoregressive coefficient and innovation variance from fitting an AR(1) model to the time series $\{f_{l,t}\}_{t=1}^T$, where $\{f_{l,t}\}$ is the l-th component of f_t.

With significance level set at 5% and $K = 10$ lags considered in the test, Table 2.2 shows the results. As N or T gets larger, the proportion of not rejecting the null hypothesis is getting closer to 95%.

2.7.3 Power of the diagnostic test

To see the power of our proposed diagnostic test in Section 2.6, we consider generating $\boldsymbol{\epsilon}_t$ using a vector autoregressive process of order one, defined by $\boldsymbol{\epsilon}_t = a\boldsymbol{I}_N\boldsymbol{\epsilon}_{t-1} + z_t$, where the z_t's are independent of each other and all elements in z_t are generated from the t_8 distribution. All other settings are the same as in Section 2.7.1. When $a = 0$, $\boldsymbol{\epsilon}_t$ is a white noise, and so the rejection rate of our diagnostic test should ideally be at the nominal level. As a increases, rejection rate should increase for our test, and eventually approach one.

	$\alpha \neq 0$ and $\gamma \neq 0$								
	H_0^1			H_0^2			H_0^3		
	$T = 50$	$T = 100$	$T = 150$	$T = 50$	$T = 100$	$T = 150$	$T = 50$	$T = 100$	$T = 150$
$N = 25$	93%	96%	97%	90%	92%	99%	93%	95%	99%
$N = 50$	91%	92%	99%	95%	98%	99%	90%	95%	99%
$N = 75$	95%	99%	99%	91%	99%	99%	95%	98%	99%
	$\alpha = 0$ and $\gamma \neq 0$								
	H_0^1			H_0^2			H_0^3		
	$T = 50$	$T = 100$	$T = 150$	$T = 50$	$T = 100$	$T = 150$	$T = 50$	$T = 100$	$T = 150$
$N = 25$	1%	5%	1%	90%	93%	98%	90%	93%	98%
$N = 50$	1%	2%	4%	95%	96%	98%	97%	96%	98%
$N = 75$	2%	7%	0%	98%	97%	98%	99%	98%	98%
	$\alpha \neq 0$ and $\gamma = 0$								
	H_0^1			H_0^2			H_0^3		
	$T = 50$	$T = 100$	$T = 150$	$T = 50$	$T = 100$	$T = 150$	$T = 50$	$T = 100$	$T = 150$
$N = 25$	92%	99%	99%	1%	3%	4%	94%	95%	99%
$N = 50$	92%	94%	99%	1%	1%	1%	92%	99%	99%
$N = 75$	97%	99%	99%	1%	3%	1%	97%	99%	99%

Table 2.1: Percentage of times when different spatial hypotheses in Section 2.5 are rejected using our proposed spatial test under different scenarios. Significance level is set at 5% in all cases.

	$\alpha \neq 0$ and $\gamma \neq 0$			$\alpha = 0$ and $\gamma \neq 0$			$\alpha \neq 0$ and $\gamma = 0$		
	$T = 50$	$T = 100$	$T = 150$	$T = 50$	$T = 100$	$T = 150$	$T = 50$	$T = 100$	$T = 150$
$N = 50$	87%	93%	93%	92%	91%	94%	89%	92%	93%
$N = 100$	89%	94%	94%	92%	94%	95%	89%	94%	94%
$N = 150$	92%	95%	95%	93%	94%	97%	94%	96%	96%

Table 2.2: Percentage of diagnostic tests in Section 2.6 that cannot be rejected under different scenarios, when underlying errors are white noise. Significance level is set at 5% in all cases, with $K = 10$ lags considered.

Figure 2.3 shows the power curves for different combinations of N and T. As T increases with N fixed, the power of the diagnostic test generally increases, while it decreases in general when N increases with T fixed. The test performs well in general for different combinations of (N, T), with power quickly approaches 1 as a increases from 0 to 0.5.

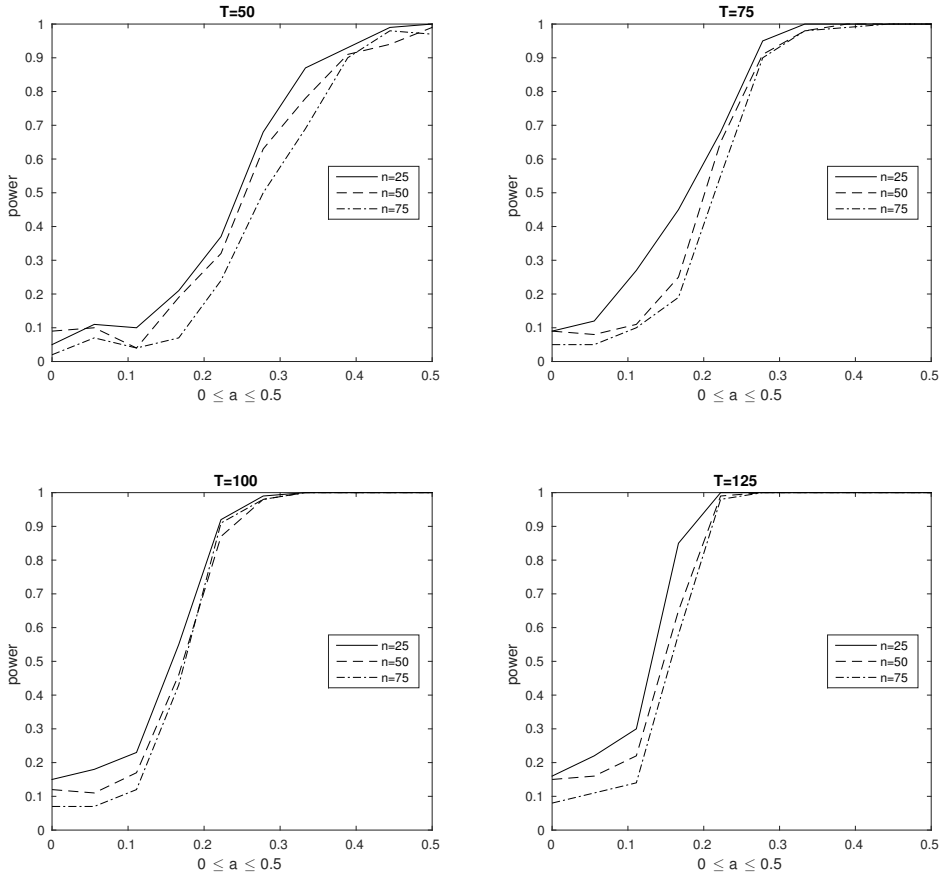


Figure 2.3: Power curves for the diagnostic test in Section 2.6 with $0 \leq a \leq 0.5$ for different (N, T) combinations. Significance level is set at 5% in all cases, with $K = 10$ lags considered.

2.7.4 Stock returns analysis

Although spatial lag models have been used extensively in analyzing economic and geographic data, financial data is rarely analyzed using such models. In fact, stock returns can be envisaged to be heavily influenced by the sector and country to which its parent company belongs. Indeed, through the use of a spatial lag model with different spatial weight matrix characterizations, Arnold et al. (2013b) shows that

stock returns belonging to the same country or same industry are spatially related to each other, with inferences on the three spatial correlation parameters in their model. Lam and Qian (2017) introduce a spatial lag model with time-lagged effects on a financial data including 32 stock prices. Their model shows the dynamic and contemporaneous spatial effects by the estimated spatial weight matrix.

In this section, we apply the quasi-maximum likelihood estimation on the same data used in Lam and Qian (2017). Based on the finding from Lam and Qian (2017), only one-lag dynamic spatial effect exists, which confirms our setting in model (2.1). However, there are some differences from Lam and Qian (2017) in this section. First, we include individual time lag effect. Then Fama-French factors mainly serving as "instrumental variables" in Lam and Qian (2017) are removed from the model. The results shown in Figure 2.4 that is similar to Figure 1.3 in Chapter 1 prove that quasi-maximum likelihood estimation when instrumental variable is not available can perform equally as least square estimation with instrumental variable. The last but not the least, inference results on these financial data are firstly introduced in this Chapter. More details can be found at the end of this section.

We aim to use our proposed model to analyze the daily log-returns of some stocks in the Euro Stoxx 50 and S&P500, providing inferences to the model parameters, and test if contemporaneous or lag-1 spatial correlations are present, while justifying our final model using our proposed diagnostic test. The stocks we used are listed below.

France	Alstom, Total, BNP, Scociete,
	Sanofi, Carrefour, LVMH, Vivendi
Germany	Daimler, Allianz, Deutsche Bank
Italy	ENEL, ENI, Intesa, Unicredit, Tele Italy
Spain	Repsol, Banco, Telefonica
US	GM, PG, Nextera, American Express,
	Citi, Wells Frago, Amgen, Gilead,
	Johnson, Costco, Home, CeNTurylink, Verizon
Energy	Alstom, Total, ENEL, ENI, Repsol, PG, Nextera
Finance	BNP, Scociete, Allianz, Deutsche Bank,
	Intesa, Unicredit, Banco, American Express,
	Citi, Wells Fargo
Pharmacy	Sanofi, Amgen, Gilead, Johnson
Retails	Carrefour, LVMH, Costco, Home
Telecom	Vivendi, Tele Italy, Telefonica, CeNTurylink, Verizon
Auto	Daimler, GM

We use the three types of spatial weight matrices used in Arnold et al. (2013b) for modeling the contemporaneous as well as the lag-1 spatial correlations in our model setting. The first spatial weight matrix W_1 used has, for $i \neq j$, the (i, j)th entry

being the weight in either the Euro Stoxx 50 or the S&P500 index for the jth stock. The second spatial weight matrix W_2 has, for $i \neq j$, the (i,j)th entry being 1 if the two stocks' parent company are in the same country. Finally, the third spatial weight matrix W_3 is similar to W_2, except that it is measuring stocks within the same industry branch instead of country. We perform row normalization on all of these three matrices.

We set two covariates, one is the national index, where we use the daily log-return of one of the five national indices S&P500, CAC40, DAX, IBEX or MIB that corresponds to where the parent company of the concerning stock belongs. The other one is the Industry index, where we use the daily log-return of the stock's corresponding industry index in Europe or US. Table 2.3 shows that all the three spatial weight

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$
Value	9.36×10^{-4}	2.2×10^{-3}	1.4×10^{-3}
S.D.	2.9×10^{-7}	4.9×10^{-5}	1.4×10^{-6}
	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$
Value	-1.8×10^{-3}	7.55×10^{-4}	1.8×10^{-3}
S.D.	4.5×10^{-7}	8.1×10^{-7}	5.7×10^{-5}
	$\hat{\phi}$	$\hat{\beta}_1$ (National Index)	$\hat{\beta}_2$ (Industry Index)
Value	1.2×10^{-3}	1.8×10^{-1}	6.9×10^{-1}
S.D.	9.4×10^{-7}	5.4×10^{-7}	4.7×10^{-6}

Table 2.3: The values of $\hat{\alpha}$, $\hat{\gamma}$, $\hat{\phi}$ and $\hat{\beta}$, with estimated standard deviations (S.D).

matrices should be included in our model for both contemporaneous and lag-1 spatial correlations. The contemporaneous spatial correlation is most affected by where the parent company belongs, followed by industry branch, but the reverse is true for the lag-1 spatial correlation. The two covariates are very important too. Figure 2.4 shows the heatmaps of the two constructed spatial weight matrices $\sum_{i=1}^{3} \hat{\alpha}_i W_i$ and $\sum_{i=1}^{3} \hat{\gamma}_i W_i$. For the former, it is obvious that there are some block patterns. These blocks represent the stocks in the same country or in the same industry. Meanwhile, the stocks are related strongly with each other if they are either all from Europe or from US. It is interesting that the ninth stock and twentieth stock are related with each other, although they belong to Germany and US auto industry respectively. The ninth stock is Daimler, while the twentieth stock is GM. But Daimler owns part of GM by spin-offs. This fact means that our linear combination of different spatial weight matrices can reflect a more general pattern and let us learn from it.

Another important feature shown in the first matrix is that there are three bars with larger values comparing with all other areas in this matrix. They represent

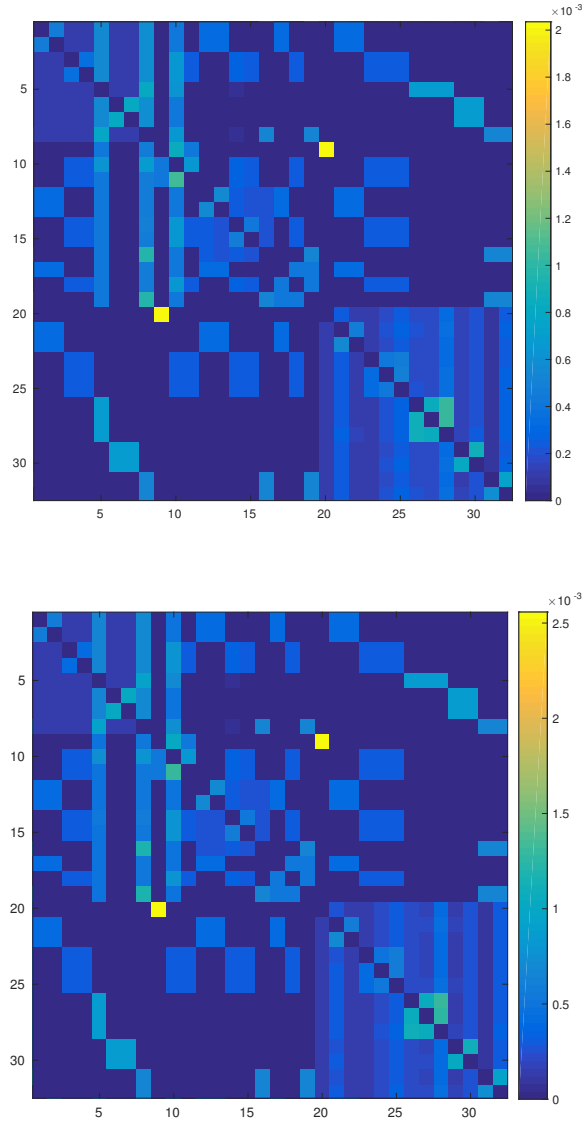


Figure 2.4: Upper: The matrix $\sum_{i=1}^{3} \hat{\alpha}_i \mathbf{W}_i$. Lower: The matrix $\sum_{i=1}^{3} \hat{\gamma}_i \mathbf{W}_i$. From 1 to 32, the stocks are Alstom, Total, BNP, Scociete, Sanofi, Carrefour, LVMH, Vivendi, Daimler, Allianz, Deutsche Bank, ENEL, ENI, Intesa, Unicredit, Tele Italy,Repsol, Banco, Telefonica, GM, PG, Nextera, American Express, Citi, Wells Frago, Amgen, Gilead, Johnson, Costco, Home, CeNTurylink and Verizon respectively.

Sanofi (France Pharmacy), Vivendi (France Telecom) and Allianz (Germany Finance). Therefore, these three stocks have a significant influence on the whole of Europe stock market. The second matrix in Figure 2.4 also shows a similar pattern.

Finally, we test for spatial correlations in the model and perform our diagnostic test. We reject all null hypotheses H_0^1 to H_0^3 in Section 2.5 due to the approximately zero p-values (4.5×10^{-8},2.3×10^{-8} and 1.0×10^{-4} respectively). Therefore, we conclude that these stock returns are under both contemporaneous spatial effects and lag-1 spatial effects. Meanwhile, using the proposed vector white noise test, we cannot reject the null hypothesis that our model residuals are white noise when we consider all lags from 1 to 10.

2.8 Conclusion

This chapter is mainly to discuss the inference for the spatial dynamic model including a Wald test on the coefficients of the linear combination of the specified spatial weight matrices and a diagnostic test whether the fitted residuals perform like a white noise vector. All theoretical results and inference are built upon the scenario when both sample size T and panel dimension N go to infinity.

The spatial dynamic model used in this chapter is a dynamic high order spatial model that the contemporaneous and time-lagged spatial effects are estimated by two linear combinations of a set of specified spatial weight matrices. Using a linear combination of spatial weight matrices not only one specific spatial weight matrix is for the avoidance of misspecification. This high order spatial model is good to avoid misspecification but more likely to include irrelevant spatial weight matrices, which makes the Wald test on the coefficients of linear combination followed more necessary. At the same time, time-lagged spatial effect is also included to figure out the delayed spatial effect. This can also be checked by the Wald test by setting a specified null hypothesis as shown in our simulation. The last but not the least, the diagnostic test on the estimated residuals is also important for the fitness of our proposed spatial dynamic model, as the fitted residuals performing like the white noise vector can prove the well fitness and correct choice on the specified spatial weight matrices of our model.

The estimation used in this chapter is quasi-maximum likelihood estimation. As shown in Theorem 2, the convergence rate $\sqrt{NT}$ comes from the fact that the fixed effects are embeded into the covariates in model (2.1). It helps to reduce the dimension of the parameter to a fixed number. Apart from this idea, we can also apply the same method used in Yu et al. (2008) and Li (2017). It is convenient to

concentrate fixed effects out by doing forward orthogonal difference or deducting the means. As the main goal here is inference not the fixed effect estimation, the proposed method makes sense here. For the future studies, there are already many existing asymptotic normality results including fixed effects such as quasi-maximum likelihood estimator in Li (2017) and generalized method of moment in Lee and Yu (2014a). We can extend our inference to these spatial dynamic model with fixed effects.

Talking about the forecasting power of the proposed model, it is same as the discussion in Chapter 1 Conclusion Section that it is hard to use our model to forecast y_t due to the large N. As the main concerns for this chapter are finding spatial weight matrix by quasi-maximum likelihood estimation and inference about the estimated coefficients of the high order spatial model and the whiteness of fitted residuals, we can leave the prediction of the proposed vector autoregression model in the future work.

2.9 Discussion for Methodologies in Chapter 1 and Chapter 2

Both model (1.1) in Chapter 1 and model (2.1) in Chapter 2 belong to the high order spatial autoregression model. It means that spatial effects are included into the model by a linear combination of pre-specified spatial weight matrices. The parameters for estimation are the coefficients in the linear combination. The main difference here is that a least square type estimation is used in Chapter 1 and the quasi-maximum likelihood estimation is applied in Chapter 2. Although the advantages and disadvantages for these two methods and relevant references have been introduced in Introduction of both Chapters, I conclude here based on the results in Chapter 1 and Chapter 2.

First, as the innate endogeneity in the spatial autoregression model, least square type estimation ,like the profiled least square estimation, only works with the existence of instrumental variables. In reality, instrumental variables are not always available. Therefore, we propose the quasi-maximum likelihood estimation in second Chapter without the usage of instrumental variable. The real data results of first two Chapters are similar, especially the estimated spatial weight matrices have an identical structure. As discussed in Section 2.7.4, the quasi-maximum likelihood estimation performs as well as the least square estimator with instrumental variable.

Second, quasi-maximum likelihood estimation is also challenged by some references due to the flat likelihood in high order spatial autoregression model. In our simulation and real data analysis, quasi-maximum likelihood estimation is computationally infeasible when M (the order) or p (the length of lag) are slightly large. To improve the performance, we find that using least square type estimator even without appropriate instrumental variables as initial value in algorithm of maximum likelihood estimation can significantly improve the validation of quasi-maximum likelihood estimation.

In conclude, both least square estimation and maximum likelihood estimation in spatial autoregression model have their pros and cons. Especially, in reality, only by using one of them can not provide a convincing result due to the lack of working-well instrumental variables and complex parameter space respectively. So it is wise to combine two methods through using maximum likelihood estimation with initial values from least square estimation when the instrumental variables are not perfectly exogenous.

2.10 Technical Proofs

To prove the asymptotic results of the quasi-maximum likelihood estimation $\hat{\theta}_{NT}$, we need to deal with the log likelihood function and it's derivatives. Note that the only stochastic part of the log likelihood function (2.3) is a function of $\boldsymbol{\epsilon}_t(\xi)$.

According to (2.1) and (2.4), we have

$$\begin{aligned}\boldsymbol{\epsilon}_t(\xi) &= y_t - \sum_{i=1}^{M} \alpha_i \mathbf{W}_{0i} y_t - \sum_{i=1}^{M} \gamma_i \mathbf{W}_{0i} y_{t-1} - \phi y_{t-1} - \mathbf{X}_t \beta \\ &= \boldsymbol{\epsilon}_t - L_{Nt}(\alpha - \alpha_0) - O_{Nt}(\delta - \delta_0). \end{aligned} \tag{2.11}$$

Due to (2.8), the stochastic part of y_t has a moving average representation and we define the following two general series to deal with it. Denote

$$\mathcal{U}_{Nt} = \sum_{h=1}^{\infty} P_{Nh} \boldsymbol{\epsilon}_{t+1-h}, \quad \mathcal{V}_{Nt} = \sum_{h=1}^{\infty} Q_{Nh} \boldsymbol{\epsilon}_{t+1-h}, \tag{2.12}$$

where P_{Nh} and Q_{Nh} are sequences of $N \times N$ non-stochastic square matrices. Before giving the proof of Theorem 1, we introduce some useful Lemmas first.

Lemma 1. *For $N \times N$ matrices $G_{1N} = (G_{1,ij})$ and $G_{2N} = (G_{2,ij})$, denote $Cov\big(\boldsymbol{\epsilon}_t' G_{1N} \boldsymbol{\epsilon}_s, \boldsymbol{\epsilon}_g' G_{2N} \boldsymbol{\epsilon}_h\big)$ by $\Delta(t,s,g,h)$. Suppose Assumption A2 holds and*

we have

$$\Delta(t,s,g,h) = (Ee_{11}^4 - 3\sigma_0^4)\sum_{i=1}^N G_{1,ii}G_{2,ii} + \sigma_0^4[tr(G_{1N}G_{2N}') + tr(G_{1N}G_{2N})],$$
$$for\ t = s = g = h,$$
$$\Delta(t,s,g,h) = \Gamma(t-g)\sum_{i=1}^N G_{1,ii}G_{2,ii},\ for\ t = s \neq g = h,$$
$$\Delta(t,s,g,h) = \Gamma(t-s)\sum_{i=1}^N G_{1,ii}G_{2,ii} + \sigma_0^4 tr(G_{1N}G_{2N}'),\ for\ t = g \neq s = h,$$
$$\Delta(t,s,g,h) = \Gamma(t-s)\sum_{i=1}^N G_{1,ii}G_{2,ii} + \sigma_0^4 tr(G_{1N}G_{2N}),\ for\ t = h \neq s = g,$$

and 0 *otherwise.*

Proof of Lemma 1.

Under Assumption A2, $\{e_{it}\}$ are uncorrelated across i and t. Thus,

$$E(\boldsymbol{\epsilon}_t' G_{1N}\boldsymbol{\epsilon}_s) = \sigma_0^2 tr(G_{1N})I(t=s), \tag{2.13}$$

where $I(t=s) = 1$ for $t = s$ and $I(t=s) = 0$ otherwise.

Note that $\{e_{it}\}$ and $\{e_{jt}\}$ are independent for $i \neq j$ under Assumption A2, and we obtain $E(e_{it}e_{js}e_{pg}e_{qh}) \neq 0$ only if $i = j$ and $p = q$, or $i = p$ and $j = q$, or $i = q$ and $j = p$. Therefore,

$$\begin{aligned}
& E\big[\sum_{i,j} e_{it}G_{1,ij}e_{js}\sum_{p,q} e_{pg}G_{1,pq}e_{qh}\big] \\
= & \sum_{i=1}^N G_{1,ii}G_{2,ii}E(e_{it}e_{is}e_{ig}e_{ih}) + \sum_{i=1}^N\sum_{p=1,p\neq i}^N G_{1,ii}G_{2,pp}E(e_{it}e_{is})E(e_{pg}e_{ph}) \\
& + \sum_{i=1}^N\sum_{j=1,j\neq i}^N G_{1,ij}G_{2,ij}E(e_{it}e_{ig})E(e_{js}e_{jh}) + \sum_{i=1}^N\sum_{j=1,j\neq i}^N G_{1,ij}G_{2,ji}E(e_{it}e_{ih})E(e_{js}e_{jg}) \\
= & \big[E(e_{it}e_{is}e_{ig}e_{ih}) - E(e_{it}e_{is})E(e_{pg}e_{ph}) - E(e_{it}e_{ig})E(e_{js}e_{jh}) - E(e_{it}e_{ih})E(e_{js}e_{jg})\big]\cdot \\
& \sum_{i=1}^N G_{1,ii}G_{2,ii} + E(e_{it}e_{is})E(e_{pg}e_{ph})tr(G_{1N})tr(G_{2N}) \\
& + E(e_{it}e_{ig})E(e_{js}e_{jh})tr(G_{1N}G_{2N}') + E(e_{it}e_{ih})E(e_{js}e_{jg})tr(G_{1N}G_{2N}').
\end{aligned}$$

Combining (2.13), we get for $t = s = g = h$

$$\begin{aligned}\Delta(t,s,g,h) &= E\big[(\boldsymbol{\epsilon}_t' G_{1N}\boldsymbol{\epsilon}_s)(\boldsymbol{\epsilon}_g' G_{2N}\boldsymbol{\epsilon}_h)\big] - E(\boldsymbol{\epsilon}_t' G_{1N}\boldsymbol{\epsilon}_s)E(\boldsymbol{\epsilon}_g' G_{2N}\boldsymbol{\epsilon}_h) \\ &= (Ee_{11}^4 - 3\sigma_0^4)\sum_{i=1}^{N} G_{1,ii}G_{2,ii} + \sigma_0^4[tr(G_{1N}G_{2N}') + tr(G_{1N}G_{2N})].\end{aligned}$$

Thus the first equality holds. Noting $E(e_{it}e_{is}) = 0$ for any $t \neq s$, we can similarly prove that the other equalities of Lemma 1 hold.

Lemma 2. *Under Assumptions A2, we have for $t \geq s$*

$$E(\mathcal{U}_{Nt}'\mathcal{V}_{Ns}) = \sigma_0^2 tr(\sum_{h=1}^{\infty} P_{N,t-s+h}'Q_{Nh}) \text{ and}$$

$$\begin{aligned}Cov(\mathcal{U}_{Nt}'\mathcal{V}_{Nt}, \mathcal{U}_{Ns}'\mathcal{V}_{Ns}) &= \sum_{h=1}^{t-s}\sum_{g=1}^{\infty}\Gamma(t-s+g-h)\sum_{i=1}^{N}(P_{Nh}'Q_{Nh})_{ii}(P_{Ng}'Q_{Ng})_{ii} \\ &+ \sum_{h=1}^{\infty}\sum_{r=1}^{\infty}\Gamma(h-r)\sum_{i=1}^{N}(P_{N,t-s+h}'Q_{N,t-s+h})_{ii}(P_{Nr}'Q_{Nr})_{ii} \\ &+ \sum_{h=1}^{\infty}\sum_{g=1}^{\infty}\Big\{\Gamma(h-g)\sum_{i=1}^{N}(P_{N,t-s+h}'Q_{N,t-s+g})_{ii}(P_{Nh}'Q_{Ng})_{ii} \\ &\qquad + \sigma_0^4 tr(P_{N,t-s+h}'Q_{N,t-s+g}Q_{Ng}'P_{Nh})\Big\} \\ &+ \sum_{h=1}^{\infty}\sum_{g=1}^{\infty}\Big\{\Gamma(h-g)\sum_{i=1}^{N}(P_{N,t-s+h}'Q_{N,t-s+g})_{ii}(P_{Ng}'Q_{Nh})_{ii} \\ &\qquad + \sigma_0^4 tr(P_{N,t-s+h}'Q_{N,t-s+g}Q_{Ng}'P_{Nh})\Big\} \\ &- 2Ee_{11}^4\sum_{h=1}^{\infty}\sum_{i=1}^{N}(P_{N,t-s+h}'Q_{N,t-s+h})_{ii}(P_{Nh}'Q_{Nh})_{ii}\end{aligned}$$

Proof of Lemma 2.

(i) By (2.13), we have

$$E(\mathcal{U}_{Nt}'\mathcal{V}_{Ns}) = \sum_{h=1}^{\infty}\sum_{r=1}^{\infty}E(\boldsymbol{\epsilon}_{t+1-h}'P_{Nh}'Q_{Nr}\boldsymbol{\epsilon}_{s+1-r}) = \sigma_0^2 tr(\sum_{h=1}^{\infty}P_{N,t-s+h}'Q_{Nh}).$$

(ii) For $t \geq s$, $\mathcal{U}_{Nt}$ and $\mathcal{V}_{Nt}$ can be rewritten as

$$\mathcal{U}_{Nt} = \sum_{h=1}^{t-s} P_{Nh}\boldsymbol{\epsilon}_{t+1-h} + \sum_{g=1}^{\infty} P_{N,t-s+g}\boldsymbol{\epsilon}_{s+1-g},$$
$$\mathcal{V}_{Nt} = \sum_{h=1}^{t-s} Q_{Nh}\boldsymbol{\epsilon}_{t+1-h} + \sum_{g=1}^{\infty} Q_{N,t-s+g}\boldsymbol{\epsilon}_{s+1-g}.$$

Under Assumption A2, we have

$$Cov\big(\sum_{h=1}^{t-s} P_{Nh}\boldsymbol{\epsilon}_{t+1-h} \sum_{g=1}^{\infty} Q_{t-s+g}\boldsymbol{\epsilon}_{s+1-g}, \mathcal{U}'_{Ns}\mathcal{V}_{Ns}\big) = 0,$$
$$Cov\big(\sum_{g=1}^{\infty} P_{N,t-s+g}\boldsymbol{\epsilon}_{N,s+1-g} \sum_{h=1}^{t-s} Q_{Nh}\boldsymbol{\epsilon}_{N,t+1-h}, \mathcal{U}'_{Ns}\mathcal{V}_{Ns}\big) = 0.$$

Thus,

$$Cov(\mathcal{U}'_{Nt}\mathcal{V}_{Nt}, \mathcal{U}'_{Ns}\mathcal{V}_{Ns}) = \Delta_1 + \Delta_2,$$

where

$$\Delta_1 = Cov\big(\sum_{h=1}^{t-s} \boldsymbol{\epsilon}'_{t+1-h}P'_{Nh} \sum_{h=1}^{t-s} Q_{Nh}\boldsymbol{\epsilon}_{t+1-h},\ \sum_{g=1}^{\infty} \boldsymbol{\epsilon}'_{n,s+1-g}P'_{Ng} \sum_{g=1}^{\infty} Q_{Ng}\boldsymbol{\epsilon}_{n,s+1-g}\big),$$
$$\Delta_2 = Cov\big(\sum_{g=1}^{\infty} \boldsymbol{\epsilon}'_{s+1-g}P'_{N,t-s+g} \sum_{g=1}^{\infty} Q_{N,t-s+g}\boldsymbol{\epsilon}_{s+1-g},\ \sum_{g=1}^{\infty} \boldsymbol{\epsilon}'_{s+1-g}P'_{Ng} \sum_{g=1}^{\infty} Q_{Ng}\boldsymbol{\epsilon}_{s+1-g}\big).$$

Using Lemma 1, we obtain

$$\begin{aligned}\Delta_1 &= \sum_{h=1}^{t-s}\sum_{g=1}^{\infty} Cov\big(\boldsymbol{\epsilon}'_{t+1-h}P'_{Nh}Q_{Nh}\boldsymbol{\epsilon}_{t+1-h},\ \boldsymbol{\epsilon}'_{s+1-g}P'_{Ng}Q_{Ng}\boldsymbol{\epsilon}_{s+1-g}\big) \\ &= \sum_{h=1}^{t-s}\sum_{g=1}^{\infty} \Gamma(t-s+g-h) \sum_{i=1}^{N} (P'_{Nh}Q_{Nh})_{ii}(P'_{Ng}Q_{Ng})_{ii},\end{aligned}$$

where and in the following, $(M)_{ii}$ denotes the (i,i)th element for any matrix M.

Let's turn to Δ_2 then. By Lemma 1, we have

$$\begin{aligned}
\Delta_2 &= \sum_{h=1}^{\infty}\sum_{g=1}^{\infty}\sum_{r=1}^{\infty}\sum_{k=1}^{\infty} Cov\big(\boldsymbol{\epsilon}'_{s+1-h}P'_{N,t-s+h}Q_{N,t-s+g}\boldsymbol{\epsilon}_{s+1-g},\ \boldsymbol{\epsilon}'_{s+1-r}P'_{Nr}Q_{Nk}\boldsymbol{\epsilon}_{s+1-k}\big) \\
&= \sum_{h=1}^{\infty} Cov\big(\boldsymbol{\epsilon}'_{s+1-h}P'_{N,t-s+h}Q_{N,t-s+h}\boldsymbol{\epsilon}_{s+1-h},\ \boldsymbol{\epsilon}'_{s+1-h}P'_{Nh}Q_{Nh}\boldsymbol{\epsilon}_{s+1-h}\big) \\
&\quad + \sum_{h=1}^{\infty}\sum_{r=1,r\neq h}^{\infty} Cov\big(\boldsymbol{\epsilon}'_{s+1-h}P'_{N,t-s+h}Q_{N,t-s+h}\boldsymbol{\epsilon}_{s+1-h},\ \boldsymbol{\epsilon}'_{s+1-r}P'_{Nr}Q_{Nr}\boldsymbol{\epsilon}_{s+1-r}\big) \\
&\quad + \sum_{h=1}^{\infty}\sum_{g=1,g\neq h}^{\infty} Cov\big(\boldsymbol{\epsilon}'_{s+1-h}P'_{N,t-s+h}Q_{N,t-s+g}\boldsymbol{\epsilon}_{s+1-g},\ \boldsymbol{\epsilon}'_{s+1-h}P'_{Nh}Q_{Ng}\boldsymbol{\epsilon}_{s+1-g}\big) \\
&\quad + \sum_{h=1}^{\infty}\sum_{g=1,r\neq h}^{\infty} Cov\big(\boldsymbol{\epsilon}'_{s+1-h}P'_{N,t-s+h}Q_{N,t-s+g}\boldsymbol{\epsilon}_{s+1-g},\ \boldsymbol{\epsilon}'_{s+1-g}P'_{Ng}Q_{Nh}\boldsymbol{\epsilon}_{s+1-h}\big) \\
&=: \Delta_{21}+\Delta_{22}+\Delta_{23}+\Delta_{24},
\end{aligned}$$

where

$$\begin{aligned}
\Delta_{21} &= (Ee_{11}^4-3\sigma_0^4)\sum_{h=1}^{\infty}\sum_{i=1}^{N}(P'_{N,t-s+h}Q_{N,t-s+h})_{ii}(P'_{Nh}Q_{Nh})_{ii} \\
&\quad +\sigma_0^4\sum_{h=1}^{\infty}\big[tr(P'_{N,t-s+h}Q_{N,t-s+h}Q'_{Nh}P_{Nh})+tr(P'_{N,t-s+h}Q_{N,t-s+h}P'_{Nh}Q_{Nh})\big], \\
\Delta_{22} &= \sum_{h=1}^{\infty}\sum_{r=1}^{\infty}\Gamma(h-r)\sum_{i=1}^{N}(P'_{N,t-s+h}Q_{N,t-s+h})_{ii}(P'_{Nr}Q_{Nr})_{ii} \\
&\quad -\Gamma(0)\sum_{h=1}^{\infty}\sum_{i=1}^{N}(P'_{N,t-s+h}Q_{N,t-s+h})_{ii}(P'_{Nh}Q_{Nh})_{ii}, \\
\Delta_{23} &= \sum_{h=1}^{\infty}\sum_{g=1}^{\infty}\Big\{\Gamma(h-g)\sum_{i=1}^{N}(P'_{N,t-s+h}Q_{N,t-s+g})_{ii}(P'_{Nh}Q_{Ng})_{ii} \\
&\qquad\qquad +\sigma_0^4 tr(P'_{N,t-s+h}Q_{N,t-s+g}Q'_{Ng}P_{Nh})\Big\} \\
&\quad -\sum_{h=1}^{\infty}\Big\{\Gamma(0)\sum_{i=1}^{N}(P'_{N,t-s+h}Q_{N,t-s+h})_{ii}(P'_{Nh}Q_{Nh})_{ii} \\
&\qquad\qquad +\sigma_0^4 tr(P'_{N,t-s+h}Q_{N,t-s+h}Q'_{Nh}P_{Nh})\Big\}
\end{aligned}$$

$$\Delta_{24} = \sum_{h=1}^{\infty}\sum_{g=1}^{\infty}\Big\{\Gamma(h-g)\sum_{i=1}^{N}(P'_{N,t-s+h}Q_{N,t-s+g})_{ii}(P'_{Ng}Q_{Nh})_{ii} + \sigma_0^4 tr(P'_{N,t-s+h}Q_{N,t-s+g}Q'_{Ng}P_{Nh})\Big\}$$
$$- \sum_{h=1}^{\infty}\Big\{\Gamma(0)\sum_{i=1}^{N}(P'_{N,t-s+h}Q_{N,t-s+h})_{ii}(P'_{Nh}Q_{Nh})_{ii} + \sigma_0^4 tr(P'_{n,t-s+h}Q_{n,t-s+h}P'_{Nh}Q_{Nh})\Big\}.$$

Therefore, we obtain the results of Lemma 2.

Lemma 3. *Suppose assumptions A1-A7 hold. Then for any $N \times N$ non-stochastic matrix $G_N = (G_{ij})$ with finite L_1 norm and L_∞ norm, it follows that*

(i) $\frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{\epsilon}'_t G_N \boldsymbol{\epsilon}_t - E\frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{\epsilon}'_t G_N \boldsymbol{\epsilon}_t = O_p(\frac{1}{\sqrt{NT}})$
and $E\frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{\epsilon}'_t G_N \boldsymbol{\epsilon}_t = O(1)$

(ii) $\frac{1}{NT}\sum_{t=1}^{T}L'_{Nt} G_N L_{Nt} - E\frac{1}{NT}\sum_{t=1}^{T}L'_{Nt} G_N L_{Nt} = O_p(\frac{1}{\sqrt{NT}})$
and $E\frac{1}{NT}\sum_{t=1}^{T}L'_{Nt} G_N L_{Nt} = O(1)$

(iii) $\frac{1}{NT}\sum_{t=1}^{T}O'_{Nt} G_N O_{Nt} - E\frac{1}{NT}\sum_{t=1}^{T}O'_{Nt} G_N O_{Nt} = O_p(\frac{1}{\sqrt{NT}})$
and $E\frac{1}{NT}\sum_{t=1}^{T}O'_{Nt} G_N O_{Nt} = O(1)$

(iv) $\frac{1}{NT}\sum_{t=1}^{T}L'_{Nt} G_N \boldsymbol{\epsilon}_t - E\frac{1}{NT}\sum_{t=1}^{T}L'_{Nt} G_N \boldsymbol{\epsilon}_t = O_p(\frac{1}{\sqrt{NT}})$
and $E\frac{1}{NT}\sum_{t=1}^{T}L'_{Nt} G_N \boldsymbol{\epsilon}_t = O(1)$

(v) $\frac{1}{NT}\sum_{t=1}^{T}O'_{Nt} G_N \boldsymbol{\epsilon}_t = O_p(\frac{1}{\sqrt{NT}})$
and $E\frac{1}{NT}\sum_{t=1}^{T}O'_{Nt} G_N \boldsymbol{\epsilon}_t = 0.$

Proof of Lemma 3.

(i) By Lemma 1, we have

$$\begin{aligned}\bar{\Delta}_1 &= \frac{1}{NT}\sum_{t=1}^{T} Var\Big(\boldsymbol{\epsilon}_t' G_N \boldsymbol{\epsilon}_t\Big) \\ &= (Ee_{11}^4 - 3\sigma_0^4)\frac{1}{N}\sum_{i=1}^{N} G_{ii}^2 + \frac{\sigma_0^4}{N}[tr(G_N G_N') + tr(G_N^2)] \\ &= O(1)\end{aligned}$$

due to the finity of $\|G_N\|_1$ and $\|G_N\|_\infty$. By Assumption A2(ii) and Lemma 1, we have

$$\begin{aligned}\bar{\Delta}_2 &= \frac{2}{NT}\sum_{s<t} Cov(\boldsymbol{\epsilon}_t' G_N \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_s' G_N \boldsymbol{\epsilon}_s) \\ &= \frac{2}{NT}\sum_{s<t} \Gamma(t-s)\sum_{i=1}^{N} G_{ii}^2 \\ &= O(1).\end{aligned}$$

Hence,

$$E\Big(\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\boldsymbol{\epsilon}_t' G_N \boldsymbol{\epsilon}_t - E\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\boldsymbol{\epsilon}_t' G_N \boldsymbol{\epsilon}_t\Big)^2 = \bar{\Delta}_1 + \bar{\Delta}_2 = O(1),$$

which implies the first part of (i) in Lemma 3. Since $\|G_N\|_1$ and $\|G_N\|_\infty$ are both finite, we have

$$E\frac{1}{NT}\sum_{t=1}^{T}\boldsymbol{\epsilon}_t' G_N \boldsymbol{\epsilon}_t = E\frac{\sigma_0^2}{N}tr(G_N) = O(1).$$

(ii) Note that the (i,j) element of $L_{Nt}' G_N L_{Nt}$ is $y_t' \mathbf{W}_{0i}' G_N \mathbf{W}_{0j} y_t$. By (2.8), we get

$$\begin{aligned}y_t' \mathbf{W}_{0i}' G_N \mathbf{W}_{0j} y_t = \quad & \mathcal{E}_{Nt}' \mathbf{W}_{0i}' G_N \mathbf{W}_{0j} \mathcal{E}_{Nt} + \mathcal{X}_{Nt}'(\mathbf{W}_{0i}' G_N \mathbf{W}_{0j} + \mathbf{W}_{nj}' G_N \mathbf{W}_{0i})\mathcal{E}_{Nt} \\ & + \mathcal{X}_{Nt}' \mathbf{W}_{0i}' G_N \mathbf{W}_{0j} \mathcal{X}_{Nt} \\ =: \quad & H_{1,Nt} + H_{2,Nt} + H_{3,Nt}.\end{aligned}$$

To prove the first part of Lemma 3, we only need to prove for $i = 1, 2, 3$,

$$\frac{1}{NT}\sum_{t=1}^{T} H_{i,NT} - E\frac{1}{NT}\sum_{t=1}^{T} H_{i,NT} = O(\frac{1}{\sqrt{NT}}), \tag{2.14}$$

Since $\{\mathcal{X}'_{Nt}\}$ are non-stochastic, it is obvious that the left side of (2.14) is zero for $i = 3$. Using lemma 8 of Yu et al. (2008), we can obtain (2.14) holds for $i = 2$.

Next, we deal with the case $i = 1$. By (2.8) and (2.12), we have $H_{1,Nt} = \mathcal{U}'_{Nt}\mathcal{V}_{Nt}$ with $P_{Nh} = \mathbf{W}_{0i}A_N^h H_N^{-1}$ and $Q_{Nh} = G_N\mathbf{W}_{0j}A_N^h H_N^{-1}$. Therefore, we only need to prove that

$$E\Big\{\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}H_{1,Nt} - E\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}H_{1,Nt}\Big\}^2 = \frac{1}{NT}Var(\sum_{t=1}^{T}\mathcal{U}'_{Nt}\mathcal{V}_{Nt}) = O(1). \quad (2.15)$$

On the other hand, by Lemma 2,

$$\begin{aligned}
&\frac{1}{NT}Var(\sum_{t=1}^{T}\mathcal{U}'_{Nt}\mathcal{V}_{Nt}) = \frac{1}{NT}\sum_{t=1}^{T}\sum_{s=1}^{T}Cov(\mathcal{U}'_{Nt}\mathcal{V}_{Nt},\ \mathcal{U}'_{Ns}\mathcal{V}_{Ns}) \\
=:\ &\frac{1}{NT}\sum_{t=1}^{T}\sum_{s=1}^{T}\big(\mathcal{D}_1(t-s) + \mathcal{D}_2(t-s) + \mathcal{D}_3(t-s) + \mathcal{D}_4(t-s) + \mathcal{D}_5(t-s)\big),
\end{aligned}$$

where $\mathcal{D}_i(t-s)$ is the corresponding part of $Cov(\mathcal{U}'_{Nt}\mathcal{V}_{Nt},\ \mathcal{U}'_{Ns}\mathcal{V}_{Ns})$ in Lemma 2. Since $\Gamma(h)$ is summable, under the Assumptions A2-A7, the same proof of Lemma 6 of Yu et al. (2008) yields that

$$\frac{1}{NT}\sum_{t=1}^{T}\sum_{s=1}^{T}\big(\mathcal{D}_2(t-s) + \mathcal{D}_3(t-s) + \mathcal{D}_4(t-s) + \mathcal{D}_5(t-s)\big) = O(1). \quad (2.16)$$

From Lemma 2, we have

$$\begin{aligned}
&\Big|\frac{1}{NT}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathcal{D}_1(t-s)\Big| \\
=\ &\Big|\frac{1}{NT}\sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{h=1}^{t-s}\sum_{g=1}^{\infty}\Gamma(t-s+g-h)\sum_{i=1}^{N}(P'_{Nh}Q_{Nh})_{ii}(P'_{Ng}Q_{Ng})_{ii}\Big| \\
\le\ &\frac{1}{N}\sum_{h=1}^{\infty}\sum_{g=1}^{\infty}\sum_{i=1}^{N}abs(P'_{Nh}Q_{Nh})_{ii}abs(P'_{Ng}Q_{Ng})_{ii}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}|\Gamma(t-s+g-h)| \\
\le\ &\frac{1}{N}\sum_{h=1}^{\infty}\sum_{g=1}^{\infty}\sum_{i=1}^{N}abs(P'_{Nh}Q_{Nh})_{ii}abs(P'_{Ng}Q_{Ng})_{ii}\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{\infty}|\Gamma(j)| \\
=\ &\sum_{j=1}^{\infty}|\Gamma(j)|\frac{1}{N}\sum_{h=1}^{\infty}\sum_{g=1}^{\infty}\sum_{i=1}^{N}abs(P'_{Nh}Q_{Nh})_{ii}abs(P'_{Ng}Q_{Ng})_{ii} \\
=\ &O(1),
\end{aligned}$$

due to Assumption A2 and the same proof of Lemma 6 of Yu et al. (2008). Combining with (2.16), we have showed that (2.14) holds for $i = 1$ and thus the first part of (ii) of Lemma 3. By Lemma 2 and similar method, we can easily show that the second part of (ii) of Lemma 3 holds.

(iii) Note that $\mathbf{X}_t$ is non-stochastic and the same method as (ii) can yield the results.

(iv) By (2.1), we have

$$\boldsymbol{\epsilon}_t = y_t - \big(\sum_{i=1}^{M} \alpha_{0i}\mathbf{W}_{0i}\big)y_t - \big(\sum_{i=1}^{M} \gamma_{0i}\mathbf{W}_{0i}\big)y_{t-1} - \phi_0 y_{t-1} + \mathbf{X}_t\beta_0.$$

Using the same method as (ii), we can get the results.

(v) By Assumption A2 and the definition of O_{Nt}, we can easily verify

$$E\frac{1}{NT}\sum_{t=1}^{T} O'_{Nt}G_N\boldsymbol{\epsilon}_t = 0.$$

Using the same method as (ii), we can obtain the results.

Lemma 4. *Suppose $U_i = (u_{i,ks})$ are $N \times N$ matrices, $i = 1, \ldots, M$. It follows that*

$$U = \begin{pmatrix} tr(U_1U_1') & tr(U_1U_2') & \cdots & tr(U_1U_M') \\ tr(U_2U_1') & tr(U_2U_2') & \cdots & tr(U_2U_M') \\ \vdots & \vdots & \vdots & \vdots \\ tr(U_MU_1') & tr(U_MU_2') & \cdots & tr(U_MU_M') \end{pmatrix}$$

is non-negative.

Proof of Lemma 4.

We only need to prove $x'Ux \geq 0$ for any M dimensional vector $x = (x_1, \ldots, x_M)'$. Note that

$$\begin{aligned} x'Ux &= \sum_{i,j} x_i tr(U_iU_j')x_j = \sum_{i,j} x_i \sum_{k,s} u_{i,ks}u_{j,ks}x_j = \sum_{k,s}\sum_{i,j} u_{i,ks}x_i x_j u_{j,ks} \\ &= \sum_{k,s}(\sum_{i} x_i u_{i,ks})^2 \geq 0. \end{aligned}$$

Thus, we complete the proof of Lemma 4.

Lemma 5. *Let Θ be any compact parameter space. Then under assumptions A1-A7, it follows that*

(i) $l_{NT}(\theta) - El_{NT}(\theta) \xrightarrow{P} 0$ uniformly in Θ,

(ii) $El_{NT}(\theta)$ is uniformly equicontinuous for $\theta \in \Theta$.

Proof of Lemma 5.

(i) By (2.5), we have

$$El_{NT}(\theta) = \frac{1}{2}\log 2\pi + \frac{1}{2}\log\sigma^2 - \frac{1}{N}\log|H_N(\alpha)| + \frac{1}{2\sigma^2 NT} E\sum_{t=1}^{T} \boldsymbol{\epsilon}_t'(\xi)\boldsymbol{\epsilon}_t(\xi). \quad (2.17)$$

From (2.5) and (2.17), we get

$$l_{NT}(\theta) - El_{NT}(\theta) = -\frac{1}{2\sigma^2}\big[\frac{1}{NT}\sum_{t=1}^{T} \boldsymbol{\epsilon}_t'(\xi)\boldsymbol{\epsilon}_t(\xi) - \frac{1}{NT} E\sum_{t=1}^{T} \boldsymbol{\epsilon}_t'(\xi)\boldsymbol{\epsilon}_t(\xi)\big].$$

The compactness of Θ implies all parameters are bounded, therefore we only need to prove

$$\frac{1}{NT}\sum_{t=1}^{T} \boldsymbol{\epsilon}_t'(\xi)\boldsymbol{\epsilon}_t(\xi) - \frac{1}{NT} E\sum_{t=1}^{T} \boldsymbol{\epsilon}_t'(\xi)\boldsymbol{\epsilon}_t(\xi) \xrightarrow{P} 0. \quad (2.18)$$

From (2.1) and (2.4), we have

$$\begin{aligned} \boldsymbol{\epsilon}_t(\xi) &= y_t - \big(\sum_{i=1}^{M} \alpha_i \mathbf{W}_{0i}\big) y_t - \sum_{i=1}^{M} \gamma_i \mathbf{W}_{0i} y_{t-1} - \phi y_{t-1} - \mathbf{X}_t\beta \\ &= L_{Nt}\alpha_0 + O_{Nt}\delta_0 + \boldsymbol{\epsilon}_t - L_{Nt}\alpha - O_{Nt}\delta \\ &= \boldsymbol{\epsilon}_t - L_{Nt}(\alpha - \alpha_0) - O_{Nt}(\delta - \delta_0). \end{aligned}$$

Therefore,

$$\begin{aligned} &\boldsymbol{\epsilon}_t'(\xi)\boldsymbol{\epsilon}_t(\xi) = \boldsymbol{\epsilon}_t'\boldsymbol{\epsilon}_t + (\alpha - \alpha_0)' L_{Nt}' L_{Nt}(\alpha - \alpha_0) + (\delta - \delta_0)' O_{Nt}' O_{Nt}(\delta - \delta_0) \\ &+2(\alpha - \alpha_0)' L_{Nt}' O_{Nt}(\delta - \delta_0) - 2(\alpha - \alpha_0)' L_{Nt}'\boldsymbol{\epsilon}_t - 2(\delta - \delta_0)' O_{Nt}'\boldsymbol{\epsilon}_t. \end{aligned} \quad (2.19)$$

By Lemma 3, (2.18) holds.

(ii) Since Θ is compact, we have $\frac{1}{2}\log\sigma^2$ is equicontinuous. According to (2.17), we

only need to prove

$$\frac{1}{N}\log|H_N(\alpha)| \text{ is equicontinuous} \tag{2.20}$$

and

$$\frac{1}{2\sigma^2 NT}E\sum_{t=1}^{T}\boldsymbol{\epsilon}_t'(\xi)\boldsymbol{\epsilon}_t(\xi) \text{ is equicontinuous.} \tag{2.21}$$

For any $\alpha^{(1)}$ and $\alpha^{(2)}$, there exists α^* between $\alpha^{(1)}$ and $\alpha^{(2)}$ such that

$$\begin{aligned}&\frac{1}{N}\log|H_N(\alpha^{(1)})| - \frac{1}{N}\log|H_N(\alpha^{(2)})| \\ &= \frac{1}{N}\Big(tr\big(\mathbf{W}_{01}H_N^{-1}(\alpha^*)\big),\cdots,tr\big(\mathbf{W}_{0M}H_N^{-1}(\alpha^*)\big)\Big)(\alpha^{(2)}-\alpha^{(1)}).\end{aligned}$$

By Assumption A5, $\frac{1}{N}tr\big(\mathbf{W}_{0i}H_N^{-1}(\alpha^*)\big)$ is bounded for $i = 1,\cdots r$, which means (2.20) holds. From (2.19), Lemma 3 and the compactness of Θ, using the same method as proving (2.20), we can obtain (2.21) holds.

Lemma 6. *Suppose Assumption A1-A7 hold. If furthermore e_{it}^2 are uncorrelated across t, then it follows that*

$$\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial\theta} \xrightarrow{d} N(0,\Omega),$$

where $\Omega = \lim_{T\to\infty}\Omega_{NT}$ with Ω_{NT} defined by (2.36).

Proof of Lemma 6.

By (2.36), we know

$$Var(\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial\theta}) = \Omega_{NT}.$$

Thus, to prove Lemma 6, it suffices to show that for any nonzero $2M+\kappa_x+2$ dimensional vector $a = ((a^1)',(a^2)',(a^3)')'$, where a^1, a^2 and a^3 are M, $M+\kappa_x+1$ and 1 dimensional vectors,

$$\frac{a'\partial l_{NT}(\theta_0)/\partial\theta}{\sqrt{Var\big(a'\partial l_{NT}(\theta_0)/\partial\theta\big)}} \xrightarrow{d} N(0,1). \tag{2.22}$$

Let $\bar{B}_N = a_1 B'_{N1} + \cdots + a_M B'_{NM}$, we have

$$\begin{aligned} a'NT\frac{\partial l_{NT}(\theta_0)}{\partial \theta} &= -\frac{1}{\sigma_0^2}\sum_{t=1}^{T}\delta_0' O'_{Nt}\bar{B}'_N\boldsymbol{\epsilon}_t - \frac{1}{\sigma_0^2}\sum_{t=1}^{T}\left[\boldsymbol{\epsilon}_t'B'_N\boldsymbol{\epsilon}_t - \sigma_0^2 tr(\bar{B}_N)\right] \\ &\quad -\frac{1}{\sigma_0^2}\sum_{t=1}^{T}(a^2)'O'_{Nt}\boldsymbol{\epsilon}_t - \frac{1}{2\sigma_0^2}\sum_{t=1}^{T}a^3\left[\boldsymbol{\epsilon}_t'\boldsymbol{\epsilon}_t - N\sigma_0^2\right] \\ &=: \sum_{t=1}^{T}\Big(\mathcal{U}'_{N,t-1}\boldsymbol{\epsilon}_t + D'_{Nt}\boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_t'M_N\boldsymbol{\epsilon}_t - tr(M_N)\Big) \\ &= \sum_{t=1}^{T}\sum_{i=1}^{N}x_{Nt,i}, \end{aligned}$$

where $D_{Nt} = \bar{B}\mathbf{X}_t\beta_0 + \mathbf{X}_t(a_{2M+2},\cdots,a_{2M+\kappa_x+1})$, $M_N = \bar{B} + a_{2M+\kappa_x+2}I_N$, and $\mathcal{U}'_{Nt}$ is defined in (2.12) with $P_{Nh} = \left[\bar{B}\Upsilon(\varphi_0) + \Upsilon((a_{M+1},\cdots,a_{2M+\kappa_x+1})')\right]A_N^h H_N^{-1}$, and

$$x_{Nt,i} = (u_{i,t-1} + d_{Nti})e_{it} + m_{N,ii}(e_{it}^2 - \sigma_0^2) + 2(\sum_{j=1}^{i-1}\tilde{m}_{N,ij}e_{jt})e_{it}, \tag{2.23}$$

with $\tilde{m}_{N,ij}$ being the (i,j)th element of $\tilde{M}_N = (M_N + M'_N)/2$. $m_{N,ii}$'s are the diagonal elements of M_N Since e_{it}^2 are uncorrelated across t, by Assumption A2 and Lemma 1, we obtain that

$$\begin{aligned} \Pi_{NT} &=: Var\Big(a'NT\frac{\partial l_{NT}(\theta_0)}{\partial\theta}\Big) \\ &= T\sigma_0^4 tr(\sum_{h=1}^{\infty}P'_{Nh}P_{Nh}) + \sigma_0^2\sum_{t=1}^{T}D'_{Nt}D_{Nt} + T(Ee_{11}^4 - 3\sigma_0^4)\sum_{i=1}^{N}m_{N,ii}^2 \\ &\quad + T\sigma_0^4\big(tr(M_NM'_N) + tr(M_N^2)\big) + 2Ee_{11}^3\sum_{t=1}^{T}\sum_{i=1}^{N}d_{Nti}m_{N,ii}, \end{aligned}$$

where $m_{N,ii}$'s are the diagonal elements of M_N and d_{Nti} is the ith element of D_{Nt}. Under Assumption A2-A7 , using the same method as in Lemma 3, we can verify that

$$\Pi_{NT} = O(NT). \tag{2.24}$$

Define the $\sigma-$ field

$$\mathcal{F}_{N,t,i} = \sigma\big(e_{11}, e_{21},\cdots,e_{N1},\cdots,e_{1,t-1},\cdots,e_{N,t-1},e_{1,t},\cdots,e_{N,t}\big)$$

with $\mathcal{F}_{N,t,0} = \mathcal{F}_{N,t-1,N}$. Let $j = N(t-1)+i$ for $t = 1, \cdots, T$ and $i = 1, \cdots, N$. Assumption A2 means that $\{x_j\}_{j=1}^{Nt}$ form a martingale difference array with respect to $\{\mathcal{F}_{j-1}\}$. According to the CLT for martingale difference in Pänsler and Prucha (1997, p.235), it is sufficient to prove that for some

$$\frac{1}{\Pi_{NT}^{1+\tau/4}} \sum_{t=1}^{T} \sum_{i=1}^{N} E|x_{Nt,i}|^{2+\tau/2} \longrightarrow 0 \text{ and } \frac{1}{\Pi_{NT}} \sum_{t=1}^{T} \sum_{i=1}^{N} E(x_{Nt,i}^2 | \mathcal{F}_{n,t,i-1}) \longrightarrow 1 \quad (2.25)$$

From (2.23), C_r inequality and Holder inequality together with Assumption A2 imply that

$$\begin{aligned}
& E|x_{Nt,i}|^{2+\tau/2} \\
\leq\ & C\Big[E|(u_{i,t-1} + d_{Nti})e_{it}|^{2+\tau/2} + |m_{N,ii}|^{2+\tau/2} E|(e_{it}^2 - \sigma_0^2)|^{2+\tau/2} + 2E|(\sum_{j=1}^{i-1} \tilde{m}_{N,ij} e_{jt})e_{it}|^{2+\tau/2}\Big] \\
\leq\ & C\big[(E|u_{i,t-1}|^{4+\tau} + |d_{Nti}|^{4+\tau})E|e_{it}|^{4+\tau}\big]^{1/2} + C + 2\big[E|\sum_{j=1}^{i-1} m_{N,ij} e_{jt}|^{4+\tau} E|e_{it}|^{4+\tau}\big]^{1/2} \\
\leq\ & C + C(E|u_{i,t-1}|^{4+\tau})^{1/2} + C(\sum_{j=1}^{i-1} |\tilde{m}_{N,ij}|)^{1/2}.
\end{aligned}$$

Since Assumption A3-A6 ensure the L_1 norm and L_∞ norm of $\sum_{h=1}^{\infty} P_{Nh}$ and M_N are bounded, we have $\sum_{j=1}^{i-1} |m_{N,ij}| \leq C$ and $E|u_{i,t-1}|^{4+\tau} \leq C$ due to Assumption A2. Therefore, $E|x_{Nt,i}|^{2+\tau/2}$, which means

$$\sum_{t=1}^{T} \sum_{i=1}^{N} E|x_{Nt,i}|^{2+\tau/2} = O(NT). \quad (2.26)$$

(2.24) and (2.26) yield

$$\frac{1}{\Pi_{NT}^{1+\tau/4}} \sum_{t=1}^{T} \sum_{i=1}^{N} E|x_{Nt,i}|^{2+\tau/2} = (NT)^{-\tau/4} \longrightarrow 0.$$

Namely, the first part of (2.25) holds.

On the other hand, by (2.23) we obtain

$$\begin{aligned} E(x_{Nt,i}^2|\mathcal{F}_{N,t,i-1}) &= \sigma_0^2\big(u_{N,t-1,i}+d_{Nti}+2\sum_{j=1}^{i-1}\tilde{m}_{N,ij}e_{jt}\big)^2+(Ee_{11}^4-\sigma_0^4)m_{N,ii}^2 \\ &\quad +2Ee_{11}^3 m_{N,ii}\big(u_{N,t-1,i}+d_{Nti}+2\sum_{j=1}^{i-1}\tilde{m}_{N,ij}e_{jt}\big). \end{aligned}$$

Denote $\tilde{M}_N^-$ be the lower diagonal matrix of $\tilde{M}_N$ and $\tilde{M}_N^o$ be the diagonal matrix with the same diagonal elements as $\tilde{M}_N$, we have

$$\begin{aligned} \sum_{i=1}^N E(x_{Nt,i}^2|\mathcal{F}_{N,t,i-1}) &= \sigma_0^2\big(U_{N,t-1}+D_{Nt}+\tilde{M}_N^-\boldsymbol{\epsilon}_t\big)'\big(U_{N,t-1}+D_{Nt}+\tilde{M}_N^-\boldsymbol{\epsilon}_t\big) \\ &\quad +(Ee_{11}^4-\sigma_0^4)\sum_{i=1}^N m_{N,ii}^2+2Ee_{11}^3\tilde{M}_N^o\big(U_{N,t-1}+D_{Nt}+\tilde{M}_N^-\boldsymbol{\epsilon}_t\big). \end{aligned}$$

Lemma 3 means the second part of (2.25) holds. Thus, we complete the proof of this lemma.

Lemma 7. *Under Assumptions A2-A8, $\Sigma = \lim_{T\to\infty}\Sigma_{NT}$ is nonsingular, where Σ_{NT} is defined in (2.37).*

Proof of Lemma 7.

Denote

$$\mathcal{J}_{\alpha\alpha}=E\sum_{t=1}^T\tilde{O}_{Nt}'\tilde{O}_{Nt},\quad \mathcal{J}_{\alpha\delta}=E\sum_{t=1}^T\tilde{O}_{Nt}'O_{Nt},\quad \mathcal{J}_{\delta\delta}=E\sum_{t=1}^T O_{NT}'O_{Nt}.$$

We have

$$E\mathcal{J}_{NT}=\begin{pmatrix}\mathcal{J}_{\alpha\alpha} & \mathcal{J}_{\alpha\delta}\\ \mathcal{J}_{\alpha\delta}' & \mathcal{J}_{\delta\delta}\end{pmatrix}.$$

Assumption A8 means that

$$\mathcal{J}_{\alpha\alpha}-\mathcal{J}_{\alpha\delta}\mathcal{J}_{\delta\delta}^{-1}\mathcal{J}_{\alpha\delta}' \quad \text{is positive definite.} \tag{2.27}$$

Let $a=((a^1)',(a^2)',(a^3)')'$ be a $2M+\kappa_x+2$ dimensional vector, where a^1, a^2 and a^3 are M, $M+\kappa_x+1$ and 1 dimensional vectors. We only need to prove that $\Sigma a=0$

implies $a = 0$. By the definition of $\mathcal{J}_{NT}$ and (2.37), we obtain

$$\Sigma = \frac{1}{\sigma_0^2} \cdot \begin{pmatrix} \mathcal{J}_{\alpha\alpha} + \lim_{T\to\infty} \frac{1}{N}\big(tr(B'_{Ni}B_{Nj}) + tr(B_{Nj}B_{Ni})\big)_{M\times M} & \mathcal{J}_{\alpha\delta} & \lim_{T\to\infty} \frac{1}{N}\tilde{tr}(B_N) \\ \mathcal{J}'_{\alpha\delta} & \mathcal{J}_{\delta\delta} & 0 \\ \lim_{T\to\infty} \frac{1}{N}\tilde{tr}(B_N)' & 0 & \frac{1}{2\sigma_0^2} \end{pmatrix}.$$

Therefore, $\Sigma a = 0$ implies

$$\begin{aligned} &\Big[\mathcal{J}_{\alpha\alpha} + \lim_{T\to\infty} \frac{1}{N}\big(tr(B'_{Ni}B_{Nj}) + tr(B_{Nj}B_{Ni})\big)_{M\times M}\Big]a^1 + \mathcal{J}_{\alpha\delta}a^2 + \lim_{T\to\infty} \frac{1}{N}\tilde{tr}(B_N)a^3 = 0, \\ &a^2 = -\mathcal{J}_{\delta\delta}^{-1}\mathcal{J}'_{\alpha\delta}a^1 \\ &a^3 = -\lim_{T\to\infty} \frac{2\sigma_0^2}{n}\tilde{tr}'(B_N)a^1. \end{aligned}$$

By eliminating a^2 and a^3, the above first equation becomes

$$\begin{aligned} &\Big\{\mathcal{J}_{\alpha\alpha} - \mathcal{J}_{\alpha\delta}\mathcal{J}_{\delta\delta}^{-1}\mathcal{J}'_{\alpha\delta} \\ &+ \lim_{T\to\infty} \frac{1}{N}\Big[\big(tr(B'_{Ni}B_{Nj}) + tr(B_{Nj}B_{Ni})\big)_{M\times M} - \frac{2}{N}\tilde{tr}(B_N)\tilde{tr}'(B_N)\Big]\Big\}a^1 = 0. \end{aligned}$$

Denote $C_N = (C_{N1}, \cdots, C_{Nr})'$ with $C_{Ni} = B_{Ni} - tr(B_{Ni})I_N/N$ and we can verify that

$$tr(B'_{Ni}B_{Nj}) + tr(B_{Nj}B_{Ni}) - \frac{2}{N}\tilde{tr}(B_{Ni})\tilde{tr}(B_{Nj}) = \frac{1}{2}tr\big[(C_{Ni} + C'_{Ni})(C_{nj} + C'_{Nj})'\big].$$

By Lemma 4 and (2.27), we know that

$$\mathcal{J}_{\alpha\alpha} - \mathcal{J}_{\alpha\delta}\mathcal{J}_{\delta\delta}^{-1}\mathcal{J}'_{\alpha\delta} + \lim_{T\to\infty} \frac{1}{N}\Big[\big(tr(B'_{Ni}B_{Nj}) + tr(B_{Nj}B_{Ni})\big)_{M\times M} - \frac{2}{N}\tilde{tr}(B_N)\tilde{tr}'(B_N)\Big]$$

is positive definite, which implies that $a^1 = 0$ and hence $a = 0$. Now, we complete the proof of Lemma 7.

Lemma 8. *Suppose Assumptions A2-A8 hold. Then, for any $\theta_* \xrightarrow{P} \theta_0$, we have*

$$\frac{\partial^2 l_{NT}(\theta_*)}{\partial\theta\partial\theta'} \xrightarrow{P} \Sigma \quad \text{and} \quad \sqrt{NT}\frac{\partial l_{NT}(\theta_*)}{\partial\theta}\sqrt{NT}\frac{\partial l_{NT}(\theta_*)}{\partial\theta'} \xrightarrow{P} \Omega.$$

Proof of Lemma 8.

We only prove the first part of this lemma since the second part can be obtained using the same method. Note that

$$\frac{\partial^2 l_{NT}(\theta_*)}{\partial\theta\partial\theta'} - \Sigma = \frac{\partial^2 l_{NT}(\theta_*)}{\partial\theta\partial\theta'} - \frac{\partial^2 l_{NT}(\theta_0)}{\partial\theta\partial\theta'} + \frac{\partial^2 l_{NT}(\theta_0)}{\partial\theta\partial\theta'} - \Sigma_{NT} + \Sigma_{NT} - \Sigma.$$

Since $\Sigma = \lim_{T\to\infty}\Sigma_{NT}$, it suffices to prove that

$$\frac{\partial^2 l_{NT}(\theta_*)}{\partial\theta\partial\theta'} - \frac{\partial^2 l_{NT}(\theta_0)}{\partial\theta\partial\theta'} \to 0 \quad \text{and} \quad \frac{\partial^2 l_{NT}(\theta_0)}{\partial\theta\partial\theta'} - \Sigma_{NT} \to 0. \tag{2.28}$$

From appendix A.2, we have

$$\begin{aligned}
\frac{\partial^2 l_{NT}(\theta_*)}{\partial\alpha_i\partial\alpha_j} - \frac{\partial^2 l_{NT}(\theta_0)}{\partial\alpha_i\partial\alpha_j} &= \big(\frac{1}{\sigma_*^2} - \frac{1}{\sigma_0^2}\big)\frac{1}{NT}\sum_{t=1}^{T}(\mathbf{W}_{0i}y_t)'\mathbf{W}_{0j}y_t \\
&\quad + \Big[\frac{1}{N\sigma_*^2}tr(B_{Ni}(\alpha_*)B_{Nj}(\alpha_*)) - \frac{1}{N\sigma_0^2}tr(B_{Ni}B_{Nj})\Big] \\
&=: E_1 + E_2.
\end{aligned}$$

Since $\sigma_*^2 \to \sigma_0^2$, Lemma 3 implies $E_1 \xrightarrow{P} 0$. By Assumption A3 and A5, the L_1 and L_∞ norm of $B_{Ni}(\alpha_*)B_{Nj}(\alpha_*)$ are bounded, which implies that $\frac{1}{N}tr(B_{Ni}(\alpha_*)B_{Nj}(\alpha_*))$ is bounded. On the other hand,

$$\begin{aligned}
&\frac{1}{N\sigma_0^2}\big[tr(B_{Ni}(\alpha_*)B_{Nj}(\alpha_*)) - tr(B_{Ni}B_{Nj})\big] \\
=\ &\frac{1}{N\sigma_0^2}tr\big[(B_{Ni}(\alpha_*) - B_{Ni})B_{Nj}(\alpha_*) + B_{Ni}(B_{Nj}(\alpha_*) - B_{Nj})\big] \\
=\ &\frac{1}{N\sigma_0^2}tr\big[\sum_{k=1}^{M} B_{Ni}(\tilde{\alpha})B_{nk}(\tilde{\alpha})(\alpha_{*k} - \alpha_{0k})B_{Nj}(\alpha_*) \\
&\qquad + B_{Ni}(\alpha_*)\sum_{k=1}^{M} B_{Nj}(\bar{\alpha})B_{nk}(\bar{\alpha})(\alpha_{*k} - \alpha_{0k})\big] \\
=\ &o_p(1),
\end{aligned}$$

where $\tilde{\alpha}$ and $\bar{\alpha}$ lie between α_* and α_0, and α_{*k} and α_{0k} are the kth element of α_* and α_0 . Thus,

$$\begin{aligned}
E_2 &= \big(\frac{1}{\sigma_*^2} - \frac{1}{\sigma_0^2}\big)\frac{1}{N}tr(B_{Ni}(\alpha_*)B_{Nj}(\alpha_*)) + \frac{1}{N\sigma_0^2}\big[tr(B_{Ni}(\alpha_*)B_{Nj}(\alpha_*)) - tr(B_{Ni}B_{Nj})\big] \\
&= o_p(1)
\end{aligned}$$

Now we have showed that

$$\frac{\partial^2 l_{NT}(\theta_*)}{\partial\alpha\partial\alpha'} - \frac{\partial^2 l_{NT}(\theta_0)}{\partial\alpha\partial\alpha'} \to 0.$$

Using the same method, we can show the first part of (2.28) holds. Lemma 3 and the expression of $\frac{\partial^2 l_{NT}(\theta_0)}{\partial\alpha\partial\alpha'}$ in appendix imply that the second part of (2.28) holds. Then we complete the proof.

Proof of Theorem 1: By (2.2) and (2.4), $\boldsymbol{\epsilon}_t(\xi)$ can be rewritten as

$$\boldsymbol{\epsilon}_t(\xi) = H_N(\alpha)H_N^{-1}O_{Nt}\delta_0 - O_{Nt}\delta + H_N(\alpha)H_N^{-1}\boldsymbol{\epsilon}_t.$$

Due to Lemma 3, we have

$$E\boldsymbol{\epsilon}_t'(\xi)\boldsymbol{\epsilon}_t(\xi) = E\big(H_N(\alpha)H_N^{-1}O_{Nt}\delta_0 - O_{Nt}\delta\big)'\big(H_N(\alpha)H_N^{-1}O_{Nt}\delta_0 - O_{Nt}\delta\big) + \sigma_0^2\mathcal{H}_N(\alpha),$$

where $\mathcal{H}_N(\alpha) = tr\big(H_N^{-1'}H_N'(\alpha)H_N(\alpha)H_N^{-1}\big)$.

Therefore, denote $Q_{Nt}(\xi) = H_N(\alpha)H_N^{-1}O_{Nt}\delta_0 - O_{Nt}\delta$ and (2.17) yields

$$\begin{aligned} El_{NT}(\theta) \;&= \tfrac{1}{2}\log 2\pi + \tfrac{1}{2}\log\sigma^2 - \tfrac{1}{N}\log|H_N(\alpha)| + \tfrac{\sigma_0^2}{2n\sigma^2}\mathcal{H}_N(\alpha) \\ &\quad + \tfrac{1}{2\sigma^2 NT}\textstyle\sum_{t=1}^{T} EQ_{Nt}'(\xi)Q_{Nt}(\xi). \end{aligned}$$

Hence,

$$El_{NT}(\theta) - El_{NT}(\theta_0) = S_{1N} + \frac{1}{2\sigma^2 NT}\sum_{t=1}^{T} EQ_{Nt}'(\xi)Q_{Nt}(\xi), \tag{2.29}$$

where $S_{1N} = \frac{1}{2}(\log\sigma^2 - \log\sigma_0^2) - \frac{1}{N}\big(\log|H_N(\alpha)| - \log|H_N(\alpha_0)|\big) + \frac{\sigma_0^2}{2n\sigma^2}\mathcal{H}_N(\alpha) - \frac{1}{2}$. Let $p(\alpha, x)$ be the density function of $\tilde{Y}(\alpha,\sigma) = H_N^{-1}(\alpha)\eta_N(\sigma)$, where $\eta_N(\sigma)$ is a multivariate normal variable with mean 0 and covariance $\sigma^2 I_N$. we can easily verify that the Kullback-leibler divergence of $p(\alpha,\sigma)$ from $p(\alpha_0,\sigma_0)$ is S_{1N}, which means $S_{1N} \geq 0$ and the equality holds if and only if $\alpha = \alpha_0$ and $\sigma = \sigma_0$ by the definition of Kullback-leibler divergence. On the other hand, $H_N(\alpha)H_N^{-1} = I_N + \sum_{i=1}^{M}(\alpha_{0i} - \alpha_i)\mathbf{W}_{0i}H_N^{-1}$ means that

$$Q_{Nt}(\xi) = O_{Nt}(\delta_0 - \delta) + \sum_{i=1}^{M}(\alpha_{0i} - \alpha_i)\mathbf{W}_{0i}H_N^{-1}O_{Nt}\delta_0.$$

Since $E\mathcal{J}_{NT}$ is nonsingular, we have $\frac{1}{2\sigma^2 NT}\sum_{t=1}^{T} EQ'_{Nt}(\xi)Q_{Nt}(\xi) \geq 0$ and the equality holds if and only if $\xi = \xi_0$. Thus, we have showed that $El_{NT}(\theta)$ has the unique minimum at $\theta = \theta_0$.

Combined with Lemma 5, the consistency of $\hat{\theta}_{NT}$ follows.

Proof of Theorem 2: According to the Taylor expansion, we have

$$\sqrt{NT}(\hat{\theta}_{NT} - \theta_0) = -\Big(\frac{1}{NT}\frac{\partial^2 l_{NT}(\bar{\theta}_{NT})}{\partial\theta\partial\theta'}\Big)^{-1}\Big(\frac{1}{\sqrt{NT}}\frac{\partial l_{NT}(\theta_0)}{\partial\theta}\Big).$$

By Lemma 6 and Lemma 8, we obtain the result.

Proof of Theorem 3: Lemma 8 implies that

$$\hat{\Sigma}_{NT} \xrightarrow{p} \Sigma \quad \text{and} \quad \hat{\Omega}_{NT} \xrightarrow{p} \Omega.$$

By Theorem 2, we get the result.

Proof of Theorem 4: According to the proof of Theorem 3 of Chang et al. (2017), we only need to prove

$$\frac{1}{T}\sum_{t=1}^{T}\big(\hat{\boldsymbol{\epsilon}}_t - \boldsymbol{\epsilon}_t\big)'\big(\hat{\boldsymbol{\epsilon}}_t - \boldsymbol{\epsilon}_t\big) = o_p(1).$$

By (2.19), Lemma 3 and Theorem 2, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\big(\hat{\boldsymbol{\epsilon}}_t - \boldsymbol{\epsilon}_t\big)'\big(\hat{\boldsymbol{\epsilon}}_t - \boldsymbol{\epsilon}_t\big)\\
=\ & (\hat{\alpha}_{NT} - \alpha_0)'\frac{1}{T}\sum_{t=1}^{T} L'_{Nt}L_{Nt}(\hat{\alpha}_{NT} - \alpha_0) + (\hat{\delta}_{NT} - \delta_0)'\frac{1}{T}\sum_{t=1}^{T} O'_{Nt}O_{Nt}(\hat{\delta}_{NT} - \delta_0)\\
& -2(\hat{\alpha}_{NT} - \alpha_0)'\frac{1}{T}\sum_{t=1}^{T} L'_{Nt}O_{Nt}(\hat{\delta}_{NT} - \delta_0)\\
=\ & O_p(\frac{1}{NT})O_p(N)\\
=\ & o_p(1).
\end{aligned}$$

Thus, we complete the proof of Theorem 4.

2.11 Appendix

To deal with the asymptotic distribution of $\hat{\theta}_{NT}$, we give the first-order and second-order derivatives.

2.11.1 The first order derivatives

From (2.5), we obtain the first order derivatives of $L_{Nt}(\theta)$ are

$$\frac{\partial l_{NT}(\theta)}{\partial \alpha} = -\frac{1}{NT\sigma^2}\sum_{t=1}^{T}\big[L'_{Nt}\boldsymbol{\epsilon}_t(\xi) - \tilde{tr}(B_N(\alpha))\big], \tag{2.30}$$

$$\frac{\partial l_{NT}(\theta)}{\partial \delta} = -\frac{1}{NT\sigma^2}\sum_{t=1}^{T}O'_{Nt}\boldsymbol{\epsilon}_t(\xi), \tag{2.31}$$

$$\frac{\partial l_{NT}(\theta)}{\partial \sigma^2} = -\frac{1}{2NT\sigma^4}\sum_{t=1}^{T}\big[\boldsymbol{\epsilon}'_t(\xi)\boldsymbol{\epsilon}_t(\xi) - N\sigma^2\big]. \tag{2.32}$$

where $\tilde{tr}(B_N(\alpha)) = (tr(B_{N1}(\alpha)), \cdots, tr(B_{NM}(\alpha)))'$ and $B_{Ni}(\alpha) = \mathbf{W}_{0i}H_N^{-1}(\alpha)$, $i = 1, \cdots, M$.

Note that $y_t = H_N^{-1}O_{Nt}\delta_0 + H_N^{-1}\boldsymbol{\epsilon}_t$, (2.30)-(2.7) yield

$$\begin{aligned}
&\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \alpha} \\
&= -\frac{1}{\sqrt{NT}\sigma_0^2}\sum_{t=1}^{T}\Big\{\tilde{O}'_{Nt}\boldsymbol{\epsilon}_t - \big[(B_{N1}\boldsymbol{\epsilon}_t, \cdots, B_{NM}\boldsymbol{\epsilon}_t)'\boldsymbol{\epsilon}_t - \sigma_0^2\tilde{tr}(B_N)\big]\Big\},
\end{aligned} \tag{2.33}$$

$$\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \delta} = -\frac{1}{\sqrt{NT}\sigma_0^2}\sum_{t=1}^{T}O'_{Nt}\boldsymbol{\epsilon}_t, \tag{2.34}$$

$$\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \sigma^2} = -\frac{1}{2\sqrt{NT}\sigma_0^2}\sum_{t=1}^{T}\big[\boldsymbol{\epsilon}'_t\boldsymbol{\epsilon}_t - N\sigma_0^2\big]. \tag{2.35}$$

Next, we compute the covariance matrix of $\sqrt{NT}\frac{\partial l_{Nt}(\theta_0)}{\partial \alpha}$. From (2.33)- (2.35)and

Assumption A2, we have

$$\begin{aligned}
\Omega_{NT,\alpha\alpha} &=: E\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \alpha}\cdot\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \alpha'} \\
&= \frac{1}{\sigma_0^2 NT}E\sum_{t=1}^{N}\tilde{O}'_{Nt}\tilde{O}_{Nt}+\frac{1}{N}\big(tr(B'_{Ni}B_{Nj})+tr(B'_{Ni}B'_{Nj})\big)_{M\times M} \\
&\quad +\frac{Ee_{11}^4-\sigma_0^4}{N\sigma_0^4}\big(\sum_{k=1}^{N}B_{Ni,kk}B_{Nj,kk}\big)_{M\times M} \quad , \\
\Omega_{NT,\alpha\delta} &=: E\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \alpha}\cdot\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \delta'}=\frac{1}{\sigma_0^2 NT}E\sum_{i=1}^{N}\tilde{O}'_{Nt}O_{Nt}, \\
\Omega_{NT,\alpha\sigma} &=: E\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \alpha_i}\cdot\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \sigma^2}=\frac{\tilde{tr}(B_N)}{N}+\frac{Ee_{11}^4-\sigma_0^4}{2N\sigma_0^4}\tilde{tr}(B_N), \\
\Omega_{NT,\delta\delta} &=: E\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \delta}\cdot\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \delta'}=\frac{1}{\sigma_0^2 NT}E\sum_{t=1}^{N}O'_{Nt}O_{Nt}, \\
\Omega_{NT,\delta\sigma} &=: E\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \delta}\cdot\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \sigma^2}=0, \\
\Omega_{NT,\sigma\sigma} &=: E\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \sigma^2}\cdot\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \sigma^2}=\frac{1}{2\sigma_0^4}+\frac{Ee_{11}^4-\sigma_0^4}{4\sigma_0^8}.
\end{aligned}$$

Note $E\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \alpha}=0$, we have

$$\Omega_{NT} =: Var\big(\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial \alpha}\big)=\begin{pmatrix}\Omega_{NT,\alpha\alpha} & \Omega_{NT,\alpha\delta} & \Omega_{NT,\alpha\sigma}\\ \Omega'_{NT,\alpha\delta} & \Omega_{NT,\delta\delta} & \Omega_{NT,\delta\sigma}\\ \Omega'_{NT,\alpha\sigma} & \Omega'_{NT,\delta\sigma} & \Omega_{NT,\sigma\sigma}\end{pmatrix}, \tag{2.36}$$

2.11.2 The second order derivatives

From (2.30)-(2.32), we have

$$\begin{aligned}
\frac{\partial^2 l_{NT}(\theta)}{\partial\alpha\partial\alpha'} &= \frac{1}{NT\sigma^2}\sum_{t=1}^{T}\Big[L'_{Nt}l_{NT} + \big(Tr(B_{Ni}(\alpha)B_{Nj}(\alpha))\big)_{M\times M}\Big] \\
\frac{\partial^2 l_{NT}(\theta)}{\partial\alpha\partial\delta'} &= \frac{1}{NT\sigma^2}\sum_{t=1}^{T}L'_{Nt}O_{Nt} \\
\frac{\partial^2 l_{NT}(\theta)}{\partial\alpha\partial\sigma^2} &= \frac{1}{NT\sigma^4}\sum_{t=1}^{T}L'_{Nt}\boldsymbol{\epsilon}_t(\xi) \\
\frac{\partial^2 l_{NT}(\theta)}{\partial\delta\partial\delta'} &= \frac{1}{NT\sigma^2}\sum_{t=1}^{T}O'_{Nt}O_{Nt} \\
\frac{\partial^2 l_{NT}(\theta)}{\partial\delta\partial\sigma^2} &= \frac{1}{NT\sigma^4}\sum_{t=1}^{T}O'_{Nt}\boldsymbol{\epsilon}_t(\xi) \\
\frac{\partial^2 l_{NT}(\theta)}{(\partial\sigma^2)^2} &= \frac{1}{NT\sigma^6}\sum_{t=1}^{T}\boldsymbol{\epsilon}'_t(\xi)\boldsymbol{\epsilon}_t(\xi) - \frac{1}{2\sigma^4}
\end{aligned}$$

Note that $y_t = H_N^{-1}O_{Nt}\delta_0 + H_N^{-1}\boldsymbol{\epsilon}_t$, by Assumption A2 we have

$$\begin{aligned}
\Sigma_{NT,\alpha\alpha} =: E\frac{\partial^2 l_{NT}(\theta_0)}{\partial\alpha\partial\alpha'} &= \frac{1}{TN\sigma^2}\sum_{t=1}^{T}\tilde{O}'_{Nt}\tilde{O}_{Nt} + \frac{1}{N\sigma^2}\big(tr(B'_{Ni}B_{Nj}) + tr(B_{Nj}B_{Ni})\big)_{M\times M}, \\
\Sigma_{NT,\alpha\delta} =: E\frac{\partial^2 l_{NT}(\theta_0)}{\partial\alpha\partial\delta'} &= \frac{1}{NT\sigma^2}E\sum_{t=1}^{T}\tilde{O}'_{Nt}O_{Nt} \\
\Sigma_{NT,\alpha\sigma} =: E\frac{\partial^2 l_{NT}(\theta_0)}{\partial\alpha\partial\sigma^2} &= \frac{1}{N}tr(B'_{Ni}), \\
\Sigma_{NT,\delta\delta} =: E\frac{\partial^2 l_{NT}(\theta_0)}{\partial\delta\partial\delta'} &= \frac{1}{NT\sigma^2}E\sum_{t=1}^{T}O'_{Nt}O_{Nt}, \\
\Sigma_{NT,\delta\sigma} =: E\frac{\partial^2 l_{NT}(\theta_0)}{\partial\delta\partial\sigma^2} &= 0, \\
\Sigma_{NT,\sigma\sigma} =: E\frac{\partial^2 l_{NT}(\theta_0)}{(\partial\sigma^2)^2} &= \frac{1}{2\sigma_0^4} - \frac{1}{T\sigma_0^4}.
\end{aligned}$$

Denote the information matrix by Σ_{NT} and we have

$$\Sigma_{NT} =: Var\big(\sqrt{NT}\frac{\partial l_{NT}(\theta_0)}{\partial\alpha}\big) = \begin{pmatrix} \Sigma_{NT,\alpha\alpha} & \Sigma_{NT,\alpha\delta} & \Sigma_{NT,\alpha\sigma} \\ \Sigma'_{NT,\alpha\delta} & \Sigma_{NT,\delta\delta} & \Sigma_{NT,\delta\sigma} \\ \Sigma'_{NT,\alpha\sigma} & \Sigma'_{NT,\delta\sigma} & \Sigma_{NT,\sigma\sigma} \end{pmatrix}. \qquad (2.37)$$

Furthermore, comparing with (2.36), we have

$$\Omega_{NT} = \Sigma_{NT} + \Xi_{NT} + O(\frac{1}{T}), \tag{2.38}$$

where

$$\Xi_{NT} = \frac{Ee_{11}^4 - 3\sigma_0^2}{\sigma_0^4} \begin{pmatrix} \left(\sum_{k=1}^N B_{Ni,kk} B_{Nj,kk}\right)_{M\times M} & 0_{M\times(M+\kappa_x+1)} & \frac{1}{2N\sigma_0^2}\tilde{tr}(B_N) \\ 0'_{M\times(M+\kappa_x+1)} & 0_{(M+\kappa_x+1)\times(M+\kappa_x+1)} & 0_{(M+\kappa_x+1)\times 1} \\ \frac{1}{2N\sigma_0^2}\tilde{tr}(B_N)' & 0'_{(M+\kappa_x+1)\times 1} & \frac{1}{4\sigma_0^4} \end{pmatrix}.$$

Chapter 3

Integrated Volatility Matrix Estimation with Nonparametric Eigenvalue Regularization

3.1 Introduction

For multivariate high frequency data analysis, a major problem concerned is the volatility matrix estimation on non-synchronized prices, allowing for the presence of microstructure noise. The problem can become more challenging when the matrix dimension p diverges with sample size n in the multivariate high frequency data. This chapter mainly proposes three estimators for large integrated volatility matrix by applying a nonparametric eigenvalue regularization on three existing multivariate realized volatility matrix estimators, which already perform well in fixed p setting.

It is commonly known that high frequency data has two main problems: microstructure noise and non-synchoronous trading times. Microstructure noise can come from many reasons, such as price discreteness and bid-ask spread bounce, which cause spurious variation in asset price. And the non-synchoronous problem refers to the situation that the different assets are traded at distinct times or their prices are observed at mismatched time points. Without these two problems, the realized volatility estimator that simply sums up all squared returns in a specified duration works efficiently. However, even only contamination from microstructure noise can make the realized volatility estimator undesirable due to inconsistency. As the discussion in Zhang (2006), the bias and variance of realized volatility estimator are

of same order as the sample size n under the discrete observation. As for the non-synchoronous trading times, Epps (1979) introduces Epps effects that asynchronicity causes high frequency covariance estimates to be biased towards zero, as sampling frequency increases.

Regardless of the high dimensional problem, there are some existing works for high frequency volatility estimation considering microstructure noise and non-synchoronous trading time for fixed dimension p. See for instance Aït-Sahalia et al. (2010) and Xiu (2010) for using maximum likelihood approaches, Griffin and Oomen (2011) for studying various existing estimators at the time assuming independent and identically distributed microstructure noise, while Zhang et al. (2005) uses a two-scale estimator to remove the bias and its multivariate extension can be easily applied by using the all-refresh time scheme with previous-tick times to overcome the problem of non-synchronous trading times. All-refresh times are defined as the time points where all assets are traded at least once starting from a previous all-refresh time point, while a previous-tick time for an asset is the last trading time before an all-refresh time. For other refresh-time schemes, see for example Fan et al. (2012).

Using the all-refresh times and previous-tick times, some more efficient nonparametric volatility matrix estimators are proposed in the recent literatures including a multi-scale realized volatility matrix (MSRVM) (Zhang, 2011), a kernel realized volatility matrix (KRVM) (Barndorff-Nielsen et al., 2011) and a pre-averaging realized volatility matrix (PRVM) (Christensen et al., 2010). These three different estimators are all consistent at different elementwise rates of convergence, but the number of assets p is assumed fixed in all cases. When p is growing with the sample size n, in particular when $p/n \to c > 0$, it is inevitable that their performances will be poorer since all three estimators are still basically a realized volatility matrix, albeit modified to remove the bias from microstructure noise. Kim et al. (2016) points out that all three estimators are inconsistent when $p/n \to c > 0$. Under the assumption of sparseness of the true integrated volatility matrix, the paper proposes to threshold the estimators and proves that a proper thresholding scheme results back in consistent estimators for all three methods. At the same time, Dai et al. (2017) improves on a pre-averaging volatility matrix estimator by assuming an underlying factor structure in the log-price processes, resulting in a low-rank plus sparse estimator, and proves consistency with rates of convergence spelt out under a number of scenarios. They perform well but the sparseness or factor structure is indispensable, which limits their application in practice.

In Lam and Feng (2018), a nonparametric eigenvalue regularized method is proposed to remove the severe bias in its extrem eigenvalues of the large sample covariance matrix under the framework $p/n \to c > 0$. This method needs no matrix structure

assumption and is built upon the two-scaled volatility matrix estimator whose rate of convergence is only at $n^{-1/6}$ shown in Zhang et al. (2005) for univariate case and in Lam and Feng (2018) for high dimensional scenario.

In this chapter, we propose the same nonparametric eigenvalue regularized MSRVM, KRVM and PRVM, abbreviated as NER-MSRVM, NER-KRVM and NER-PRVM respectively, as all of MSRVM, KRVM and PRVM can all obtain the best achieved rate on elementwise, which is $n^{-1/4}$ and same as that in parametric estimator for volatility, when the true process is Marcov. The multi-scaled estimator NER-MSRVM in particular is a generalization of the two-scaled estimator proposed in Lam and Feng (2018). Our contributions are three-fold. First of all, our estimators are more flexible to use since they do not need any structural assumptions on the true integrated volatility matrix or the log-price processes as opposed to those in Kim et al. (2016) and Dai et al. (2017). See Section 3.6 also where we have demonstrated empirically that our estimators outperform other state-of-the-art competitors in many scenarios.

Secondly, although without any structural assumptions we are not able to prove consistency, we can prove that all regularized estimators are positive definite in probability with the Corollary 4 in Lam (2016), and provide rate of convergence in spectral norm of each estimator to a rotation-equivariant "ideal" estimator, to be defined in Section 3.3.4, all under the framework $p/n \to c > 0$. For NER-MSRVM, it turns out that without sparsity assumption, the spectral norm rate of convergence is independent of p and is at $n^{-1/6}$ only when we modify the largest scales of the estimator from $n^{1/2}$ used in Kim et al. (2016), to $n^{2/3}$. With NER-KRVM and NER-PRVM, when using the positive definite versions as in Barndorff-Nielsen et al. (2011) and Christensen et al. (2010) respectively, their rates of convergence to the ideal estimator are both $n^{-1/5}$, while it is $n^{-1/4}$ when using their bias-corrected versions. These rates of convergence are the same as in the respective papers, but we are working in the framework $p/n \to c > 0$. More importantly, while the bias-corrected versions of KRVM and PRVM which converge at a rate of $n^{-1/4}$ are not guaranteed to be positive semi-definite, we prove in Theorem 2 and 3 that their corresponding regularized versions, NER-KRVM and NER-PRVM, are both positive definite in probability. See Section 3.4 for more details.

Finally, we show that with wavelet jumps-removal proposed in Fan and Wang (2007), all rates of convergence remain the same. We also prove that all estimators remain the convergence if the log-price processes are from a factor model with random or non-random drift. These results support that pre-averaging and kernel methods are better method among the three, since the rate of convergence is the fastest and is adaptive to our regularization method with jumps removal, without any extra modifications. It is also found that the pre-averaging estimator's practical performance

is the best (the regularized bias-corrected version) as seen in the empirical results in Section 3.6.

The rest of the chapter is organized as follows. Section 3.2 presents all necessary notations and the model used for a log-price process. Challenges for analyzing high frequency data are also explained. Section 3.3 introduces MSRVM, KRVM and PRVM. How nonlinear shrinkage works on these three estimators are detailed in Section 3.3.4. Asymptotic theories and detailed assumptions, including those with jumps-removal in the case of jump-diffusion log-price processes, can be found in Section 3.4. Practical concerns and implementation can be found in Section 3.5, while all simulations and a thorough empirical study are presented in Section 3.6. All proofs of theorems in the paper are presented in Section 3.7.

3.2 Model and Notations

3.2.1 Price model

We use $\mathbf{X}_t = (X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(p)})^T$ to denote a vector of log prices of p assets at time t. It follows a continuous-time diffusion model as

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\sigma}_t d\mathbf{W}_t, \quad t \in [0,1], \tag{3.1}$$

with $\{\mathbf{W}_t\}$ being a p-dimensional standard Brownian motion and $\boldsymbol{\mu}_t$ a càdlàg (elementwise) process, which can be random and correlated with $\{\mathbf{W}_t\}$ in general. The volatility $\boldsymbol{\sigma}_t \in \mathbb{R}^{p\times p}$ is also càdlàg.

The goal of this chapter is to propose the nonparametric efficient estimator for the integrated volatility matrix defined as

$$\boldsymbol{\Sigma}(a,b) = \int_a^b \boldsymbol{\sigma}_u \boldsymbol{\sigma}_u^{\mathrm{T}} du$$

for a period $[a,b]$ where $0 \le a \le b \le 1$,. As known, the covariance structure of assets plays a key role to the solution of fundamental economic problems, such as optimal asset allocation and risk management. For instance, if we have a portfolio $\mathbf{w}$ which stays constant over the period $[a,b]$, then the accumulated risk of the portfolio over the period is

$$R^{1/2}(\mathbf{w}) = (\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}(a,b)\mathbf{w})^{1/2} = \Big(\int_a^b \mathbf{w}^{\mathrm{T}}\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t^{\mathrm{T}}\mathbf{w}dt\Big)^{1/2},$$

where the term $\mathbf{w}^{\mathrm{T}}\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t^{\mathrm{T}}\mathbf{w}$ can be considered an instantaneous square-risk at time t. We take the large integrated volatility matrix as the target, allowing the presence of microstructure noise and asynchronicity under the framework of $p/n \to c > 0$.

3.2.2 Data Splitting

On the top, for the flow of this chapter and the notation consistency, we introduce the data splitting first, which is used in the proposed nonparametric eigenvalue regularization. Similar to Lam and Feng (2018), our regularized estimators are defined over a partition of the normalized time period $[0, 1]$. Let L be the total number of partitions, and

$$0 = \tau_0 < \tau_1 < \cdots < \tau_L = 1,$$

where $(\tau_{\ell-1}, \tau_\ell]$ denotes the ℓth partition. In brief, we estimate integrated volatility matrix for each partition with the help from the outside of this partition, then sum these L partition estimators together as the final estimator. More details are shown in Section 3.3.

3.2.3 Asynchronicity and microstructure noise

As shown in the Section 3.1, the first two challenges in our problem are asynchroncity and microstructure noise in multivariate high frequency data. To construct the integrated volatility matrix based on the non-synchronized high frequency data, we should apply a specific time scheme to synchronize the observation time points. There are mainly three synchronization schemes in the current literature including previous stick, all-refresh time and generalized sampling time. We use the all-refresh time scheme, which is also applied in Zhang (2011) and Lam and Feng (2018).

An all-refresh time is the time until all assets are traded at least once from the past all-refresh time point. Let $n(\ell)$ be the number of all-refresh time points in the partition $(\tau_{\ell-1}, \tau_\ell]$, $\ell = 1, \ldots, L$, so that the total number of all-refresh time points is nL, where $n = L^{-1}\sum_{\ell=1}^{L} n(\ell)$ is the average number of all-refresh time points in each partition. In this paper, we assume n is the same order as $n(\ell)$ for all ℓ, and is also the same order as nL since L is assumed finite throughout the chapter.

Let $\{v_s\}_{\{s=1,\ldots,nL\}}$ be the set of all-refresh times in $[0, 1]$ for $\mathbf{X}_t$. Then all estimators in this paper are calculated based on the previous-tick times $t_s^j \in (v_{s-1}, v_s]$, $s =$

$1, \ldots, nL$, $j = 1, \ldots, p$, which is the last trading time for the jth asset before time v_s. Note that $t_s^{j_1} \neq t_s^{j_2}$ for $j_1 \neq j_2$ in general.

In light of the contamination of microstructure noise, using the above notations, we assume the observed high frequency data $\mathbf{Y}(s)$ obeys the model as

$$\mathbf{Y}(s) = \mathbf{X}(s) + \boldsymbol{\epsilon}(s), \ s = 1, \ldots, nL, \tag{3.2}$$

where $\{\boldsymbol{\epsilon}_t\}$ is the process of microstructure noise, with $\mathbf{X}(s) = (X_{t_s^1}^{(1)}, \ldots, X_{t_s^p}^{(p)})^{\mathrm{T}}$ and $\boldsymbol{\epsilon}(s) = (\epsilon_{t_s^1}^{(1)}, \ldots, \epsilon_{t_s^p}^{(p)})^{\mathrm{T}}$. Note that $\boldsymbol{\epsilon}(\cdot)$ can be dependent on $\mathbf{X}(\cdot)$ in general.

3.3 Integrated Volatility Matrix Estimators

We known that the realized volatility matrix, which simply sums up all squared returns, performs poorly even only microstructure noise is considered. In practice, a popular method in Finance is to sparsely selected sample to ease the negative impact from microstructure noise. However, it is still inconsistent and too arbitrary. Zhang et al. (2005) first proposes two-scaled estimator whose multivariate extension is used in Lam and Feng (2018). This two-scaled estimator has three advantages: asymptotic unbiasedness, consistency and asymptotic normality. But its rate is not satisfactory.

Therefore, more efficient integrated volatility matrix estimators are proposed in recent studies. We first introduce three existing integrated volatility matrix estimators before presenting their regularized versions. They are the multi-scale realized volatility matrix estimator (MSRVM) (extended from the univariate version of Zhang (2006)), the kernel realized volatility matrix estimator (KRVM) by Barndorff-Nielsen et al. (2011) and the pre-averaging realized volatility matrix estimator (PRVM) by Christensen et al. (2010). They all have an elementwise rate of convergence of $n^{-1/4}$, which is faster then the rate $n^{-1/6}$ for the two-scale realized volatility matrix estimator (TSRVM) analyzed in Zhang (2011). Kim et al. (2016) shows that the positive-definite versions of KRVM and PRVM can still achieve an elementwise rate $n^{-1/5}$, which is still faster than $n^{-1/6}$.

Lam and Feng (2018) introduces the nonparametric eigenvalue regularized integrated volatility matrix estimator (NERIVE) which is based on a modification of the TSRVM. As their spectral norm rate of convergence to an "ideal" estimator (see Section 3.3.4 for more details) is only $n^{-1/6}$, the same as the elementwise rate for TSRVM, we analyze in Section 3.4 if our eigenvalue regularization can be applied on

MSRVM, KRVM and PRVM to achieve better rates of convergence and improved in practical performance. Before that, let us review MSRVM, KRVM and PRVM first.

3.3.1 Multi-scale realized volatility matrix

It is observed that combining the square returns from two time scales is better than the realized volatility estimator, which is the one-scale estimator. A natural question arises as to combine more than two time scales for a further improvement in estimator's efficiency. Zhang (2006) proposes a multi-scale approach in stochastic volatility estimation to eliminate the bias from microstructure noise. Our method bases on the multi-scale realized volatility matrix (MSRVM) analyzed in Tao et al. (2013). For $\ell = 1, \ldots, L$, define the MSRVM on each partition $(\tau_{\ell-1}, \tau_\ell]$ to be

$$\begin{aligned}
\mathbf{MS}(\mathbf{Y})_\ell &= \sum_{m=1}^{M} a_m [\mathbf{Y}, \mathbf{Y}^{\mathrm{T}}]_\ell^{(m)} + \zeta \left([\mathbf{Y}, \mathbf{Y}^{\mathrm{T}}]_\ell^{(1)} - [\mathbf{Y}, \mathbf{Y}^{\mathrm{T}}]_\ell^{(M)} \right), \text{ where} \\
[\mathbf{Y}, \mathbf{Y}^{\mathrm{T}}]_\ell^{(m)} &= \frac{1}{K_m} \sum_{s \in S^\ell(m)} (\mathbf{Y}(s) - \mathbf{Y}(s - K_m))(\mathbf{Y}(s) - \mathbf{Y}(s - K_m))^{\mathrm{T}}, \\
S^\ell(m) &= \{s : t_s^i, t_{s-K_m}^i \in (\tau_{\ell-1}, \tau_\ell] \text{ for all } i\}, \quad |S^\ell(m)|_m = \frac{|S^\ell(m)| - K_m + 1}{K_m}, \\
K_m &= N + m, \; a_m = \frac{12(m + N)(m - M/2 - 1/2)}{M(M^2 - 1)}, \quad \zeta = \frac{(M + N)(N + 1)}{(n + 1)(M - 1)},
\end{aligned} \tag{3.3}$$

with $N \asymp n^{2/3}$ and $M \asymp n^{1/2}$, where $a \asymp b$ means that $a = O(b)$ and $b = O(a)$. For the ease of presentation of our regularized estimators in Section 3.3.4, we also use the notation

$$\widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell)^M = \mathbf{MS}(\mathbf{Y})_\ell. \tag{3.4}$$

The above estimator is different from the one considered in Tao et al. (2013) in that we have set $N \asymp n^{2/3}$ rather than $n^{1/2}$ which is used in Tao et al. (2013) and Kim et al. (2016). The parameter N controls the magnitude of the scales used in the estimator. The reason that we use a different magnitude of scale is that unlike Tao et al. (2013) and Kim et al. (2016), we do not assume sparsity of the underlying integrated volatility matrix. Without this assumption, a diverging p can significantly increase the bias contributed from microstructure noise and asynchronous transactions. In the end, our proofs reveal that only a scale of magnitude $N \asymp n^{2/3}$ can remove the bias effects from a diverging p of the same order as n, even with our eigenvalue

regularization to be presented in Section 3.3.4. The final rate of convergence then goes from $n^{-1/4}$, the best univariate rate, to $n^{-1/6}$. See Theorem 1 in Section 3.4 for more details. Incidentally, this scale magnitude is what is used in Lam and Feng (2018), leading to the same rate of convergence $n^{-1/6}$ for the nonparametric eigenvalue regularized two-scale realized volatility matrix estimator.

In Section 3.6, the performance of the corresponding regularized estimator using $N \asymp n^{2/3}$ is also compared to that using $N \asymp n^{1/2}$, and it is clear that the one using $N \asymp n^{2/3}$ has a better practical performance.

Remark 6. *Two important conditions to be satisfied for the MSRVM are that* $\sum_{m=1}^{M}(a_m/K_m) = 0$ *and* $\sum_{m=1}^{M} a_m = 1$ *(Tao et al., 2013, Zhang, 2006). These conditions ensure the estimator has bias contributed from the microstructure noise removed, and is unbiased, respectively, resulting at an elementwise rate of convergence* $n^{-1/4}$*. Zhang (2006) also shows that efficiency can be improved by combining more than two time scales. This is also the reason why our regularized estimator based on MSRVM should perform better than the corresponding two-scale version, which is supported by our simulations in Section 3.6. However, there is no choice of* a_m *such that a positive semi-definite estimator is guaranteed. With our nonparametric eigenvalue regularization though, we can prove positive definiteness in probability for our regularized estimator under* $p/n \to c > 0$*. See Theorem 1 in Section 3.4 for more details.*

3.3.2 Kernel realized volatility matrix

Barndorff-Nielsen et al. (2011) proposes a multivariate realized kernel estimator that dose not smooth the covariance but the autocovariance operators. The estimator is robust to microstructure noise and can handle asynchronous trading. It also has a positive semi-definite version with a slightly different definition but a slower convergence rate.

Barndorff-Nielsen et al. (2011) indicates that averaging J prices at the very beginning and end of the period for a consistent estimator when kernel is applied. They name it as jittering, and with J going to infinity at an appropriate rate, the error at the beginning and the end of the day is averaged away. We adopt this scheme and denote the jittered log-price vector as $\mathbf{Y}_{\ell}^{(J)}(s),\ s = 0, 1, \ldots, n(\ell) - 2J + 1,\ \ell = 1, \ldots, L.$

Define

$$\mathbf{K}(\mathbf{Y})_\ell = \boldsymbol{\gamma}_\ell^{(0)}(\mathbf{Y}_\ell^{(J)}) + \sum_{h=1}^{n(\ell)-2J} k\left(\frac{h-1}{H}\right)\left[\boldsymbol{\gamma}_\ell^{(h)}(\mathbf{Y}_\ell^{(J)}) + \boldsymbol{\gamma}_\ell^{(-h)}(\mathbf{Y}_\ell^{(J)})\right], \quad \text{where}$$
$$\boldsymbol{\gamma}_\ell^{(h)}(\mathbf{Y}_\ell^{(J)}) = \sum_{s=h+1}^{n(\ell)-2J+1} \left(\mathbf{Y}_\ell^{(J)}(s) - \mathbf{Y}_\ell^{(J)}(s-1)\right)\left(\mathbf{Y}_\ell^{(J)}(s-h) - \mathbf{Y}_\ell^{(J)}(s-h-1)\right)^{\mathrm{T}}, \tag{3.5}$$

with $h \geq 0$, $k(\cdot)$ a kernel function and H a bandwidth parameter. We also define $\boldsymbol{\gamma}_\ell^{(h)}(\mathbf{Y}_\ell^{(J)}) = \boldsymbol{\gamma}_\ell^{(-h)}(\mathbf{Y}_\ell^{(J)})$ for $h < 0$, and assume that (i) $k(0) = 1$ and $k'(0) = 0$; (ii) $k(\cdot)$ is twice differentiable with continuous derivatives; (iii) $\int_0^\infty k(x)^2 dx$, $\int_0^\infty k'(x)^2 dx$ and $\int_0^\infty k''(x)^2 dx$ are finite. We also use the notation

$$\widetilde{\boldsymbol{\Sigma}}(\tau_{\ell-1}, \tau_\ell)^K = \mathbf{K}(\mathbf{Y})_\ell. \tag{3.6}$$

For the positive semi-definite version, assume further (iv)$\int_{-\infty}^\infty k(x)\exp(ix\lambda)dx \geq 0$ for all $\lambda \in \mathbb{R}$. Then the positive semi-definite KRVM for the partition $(\tau_{\ell-1}, \tau_\ell]$ is defined as

$$\widetilde{\boldsymbol{\Sigma}}(\tau_{\ell-1}, \tau_\ell)^{KP} = \boldsymbol{\gamma}_\ell^{(0)}(\mathbf{Y}_\ell^{(J)}) + \sum_{h=1}^{n(\ell)-2J} k\left(\frac{h}{H}\right)\left[\boldsymbol{\gamma}_\ell^{(h)}(\mathbf{Y}_\ell^{(J)}) + \boldsymbol{\gamma}_\ell^{(-h)}(\mathbf{Y}_\ell^{(J)})\right]. \tag{3.7}$$

3.3.3 Pre-avaraging realized volatility matrix

The idea of pre-averaging is to smooth the high frequency data first. The KRVM in (3.6) is also a kind of smoothing mechanism, but it only smooths the autocovariance operators instead of the data. It is intuitive that the form of smoothing of the observed log price should tend to diminish the impact of noise. The data smoothing idea is first presented in Hayashi et al. (2005), and Christensen et al. (2010) extends it to the case with microstructure noise by using pre-averaging and shows that the resulting estimator is consistent. This estimator has the property that it can be implemented directly on irregular, asynchronous and noisy observations without any form of imputation. Christensen et al. (2010) also discusses that apart from border terms, the pre-averaging estimator coincides with kernel-based estimator using the flat-top kernel function. However, kernel needs to apply some averaging to edge terms, while the pre-averaging estimator is asymptotically mixed normal by construction.

We adopt a different bias-correction term than Christensen et al. (2010) used, which is needed for our regularization to be introduced in Section 3.3.4 to work. Let $\mathbf{Y}_\ell(j)$ be the jth all-refresh log-price vector within the ℓth partition, $j = 1, \ldots, n(\ell)$. Then define

$$\begin{aligned}
\mathbf{P}(\mathbf{Y})_\ell &= \frac{1}{\psi Q} \sum_{j=1}^{n(\ell)-Q+1} \big[\bar{\mathbf{Y}}_j^{(\ell)} \bar{\mathbf{Y}}_j^{(\ell)\mathrm{T}} - \varsigma \hat{\boldsymbol{\eta}}^{(\ell)}\big], \quad \text{where} \\
\hat{\boldsymbol{\eta}}^{(\ell)} &= \frac{1}{2n(\ell)} \sum_{s=2}^{n(\ell)} (\mathbf{Y}_\ell(s) - \mathbf{Y}_\ell(s-1))(\mathbf{Y}_\ell(s) - \mathbf{Y}_\ell(s-1))^{\mathrm{T}}, \\
\bar{\mathbf{Y}}_j^{(\ell)} &= \sum_{l=1}^{Q-1} g\Big(\frac{l}{Q}\Big)[\mathbf{Y}_\ell(j+l) - \mathbf{Y}_\ell(j+l-1)], \\
\varsigma &= \sum_{l=0}^{Q-1} \Big[g\Big(\frac{l}{Q}\Big) - g\Big(\frac{l+1}{Q}\Big)\Big]^2, \quad \psi = \int_0^1 g(t)^2 dt,
\end{aligned} \tag{3.8}$$

with Q a bandwidth parameter of order $n^{1/2}$. The function $g(\cdot)$ is continuous and piecewise continuously differentiable with a piecewise Lipschitz derivative g', satisfying $g(0) = g(1) = 0$ and $\int_0^1 g(t)^2 dt > 0$. We also use the notation

$$\widetilde{\boldsymbol{\Sigma}}(\tau_{\ell-1}, \tau_\ell)^P = \mathbf{P}(\mathbf{Y})_\ell. \tag{3.9}$$

The diagonal elements in $\hat{\boldsymbol{\eta}}^{(\ell)}$ can certainly be replaced by those defined in Christensen et al. (2010) which will then be more accurate since it used up all available tick-by-tick data, but we do need off-diagonal elements to be non-zero as well, and $\hat{\boldsymbol{\eta}}^{(\ell)}$ defined above is sufficient to give us the best rate for our regularized estimator.

For the positive semi-definite version, with Q of order $n^{3/5}$, define

$$\widetilde{\boldsymbol{\Sigma}}(\tau_{\ell-1}, \tau_\ell)^{PP} = \frac{1}{\psi Q} \sum_{j=1}^{n(\ell)-Q+1} \bar{\mathbf{Y}}_j^{(\ell)} \bar{\mathbf{Y}}_j^{(\ell)\mathrm{T}}. \tag{3.10}$$

3.3.4 Nonparametric eigenvalue regularization

Multi-scale, kernel and pre-averaging realized volatility matrix estimators perform well in terms of removing the bias contributed from microstructure noise. Yet under the setting $p/n \to c > 0$, a typical realized volatility matrix suffers from bias in its extreme eigenvalues. Also, the spread of the eigenvalues of the estimator is typically much larger than its population counterpart. This creates instability in applications. Since we do not have any structural assumptions like sparsity of the

underlying matrix, regularization with sparsity such as those introduced in Kim et al. (2016) cannot be used.

The above methods all assume some form of independence for the data in proving asymptotic results. For high frequency data with contamination from microstructure noise, we cannot assume independence of the log-returns. A breakthrough comes from Lam and Feng (2018) which utilizes a data splitting scheme the same as in Section 3.2.2 to perform eigenvalue regularization for the TSRVM. They do not assume serial independence of microstructure noise, which can also be cross-sectionally correlated in general. All log-prices can also be correlated with the microstructure noise too. Their estimator is positive definite in probability under the setting $p/n \to c > 0$, and the spread of eigenvalues is shrunk within the spread of the population integrated volatility matrix. With these advantages, below we introduce a similar regularization scheme for the MSRVM, KRVM and PRVM. As discussed in the Introduction, we hope to achieve a rate of convergence faster than $n^{-1/6}$, the rate proved in Lam and Feng (2018).

First, using the data splitting scheme in Section 3.2.2, we introduce a rotation-equivariant estimator $\mathbf{\Sigma}(\mathbf{D}) = \mathbf{P}_{-j}\mathbf{D}\mathbf{P}_{-j}^{\mathrm{T}}$, where $\mathbf{D}$ is a diagonal matrix, and $\mathbf{P}_{-j}$ is orthogonal such that

$$\widetilde{\mathbf{\Sigma}}_{-j} = \mathbf{P}_{-j}\mathbf{D}_{-j}\mathbf{P}_{-j}^{\mathrm{T}}, \quad j = 1, \ldots, L, \quad \text{with } \widetilde{\mathbf{\Sigma}}_{-j} = \sum_{\ell \neq j} \widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell). \tag{3.11}$$

Here, $\widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell)$ can be $\widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell)^M$, $\widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell)^K$ or $\widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell)^P$ in (3.4), (3.6) and (3.9) respectively, or $\widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell)^{KP}$, $\widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell)^{PP}$ in (3.7) and (3.10) for the positive semi-definite versions of KRVM and PRVM respectively. This class of estimators is used in Ledoit and Wolf (2012), Lam (2016) and Lam and Feng (2018) as a starting point for the construction of their regularized estimators. First introduced in James and Stein (1961) for estimating a covariance matrix under the Stein's loss function, this class is invariant under rotation, and serves as a good starting point with no a priori information of the eigenvectors of the population covariance matrix.

Next, we want to bring $\widetilde{\mathbf{\Sigma}}_{-j}$ to be as close to $\mathbf{\Sigma}(\tau_{j-1}, \tau_j)$ as possible by fixing its eigenvectors but varying its eigenvalues. This essentially makes it a p-dimensional problem. To do this, we consider the following optimization problem:

$$\min_{\mathbf{D} \text{ diagonal}} \left\| \mathbf{P}_{-j}\mathbf{D}\mathbf{P}_{-j}^{\mathrm{T}} - \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \right\|_F, \tag{3.12}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. The same problem is also considered in Lam and Feng (2018). The reason to use $\mathbf{P}_{-j}$ for the rotation-equivariant class is

that we can condition all information outside partition j ($\mathbf{P}_{-j}$ is then fixed), so that then the correlation between $\{\mathbf{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$, and the serial correlation in $\{\boldsymbol{\epsilon}_t\}$ within partition j can be weakened. See Assumption (E3) in Section 3.4. If each partition is "small", then each $\mathbf{P}_{-j}$ is "close" to each other intuitively since each $\mathbf{P}_{-j}$ uses up all information except those within the partition.

The solution for (3.12) is $\mathbf{D} = \text{diag}(\mathbf{P}_{-j}^{\text{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{P}_{-j})$, where $\text{diag}(A)$ is a diagonal matrix with the diagonal elements of A. See Proposition 1 in Lam and Feng (2018) and the proof therein. It is not difficult to see that the spread of eigenvalues of such $\mathbf{D}$ is constrained within that of $\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)$. In Section 3.4, we present in Theorem 1, 2 and 3 that in spectral norm, $\text{diag}(\mathbf{P}_{-j}^{\text{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^M\mathbf{P}_{-j})$, $\text{diag}(\mathbf{P}_{-j}^{\text{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^K\mathbf{P}_{-j})$, $\text{diag}(\mathbf{P}_{-j}^{\text{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^P\mathbf{P}_{-j})$, $\text{diag}(\mathbf{P}_{-j}^{\text{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^{KP}\mathbf{P}_{-j})$ and $\text{diag}(\mathbf{P}_{-j}^{\text{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^{PP}\mathbf{P}_{-j})$ are converging to $\mathbf{D} = \text{diag}(\mathbf{P}_{-j}^{\text{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{P}_{-j})$ in probability. Therefore, our regularized volatility matrix estimators of MSRVM, KRVM and PRVM for the partition $(\tau_{j-1}, \tau_j]$ are respectively

$$\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^M = \mathbf{P}_{-j}\text{diag}(\mathbf{P}_{-j}^{\text{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^M\mathbf{P}_{-j})\mathbf{P}_{-j}^{\text{T}}, \tag{3.13}$$

$$\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^K = \mathbf{P}_{-j}\text{diag}(\mathbf{P}_{-j}^{\text{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^K\mathbf{P}_{-j})\mathbf{P}_{-j}^{\text{T}}, \tag{3.14}$$

$$\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^P = \mathbf{P}_{-j}\text{diag}(\mathbf{P}_{-j}^{\text{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^P\mathbf{P}_{-j})\mathbf{P}_{-j}^{\text{T}}. \tag{3.15}$$

We also denote $\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^{KP}$ and $\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^{PP}$ the corresponding estimators using $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^{KP}$ and $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^{PP}$ respectively. The corresponding nonparametric eigenvalue regularized integrated volatility matrix estimators for the period $[0, 1]$ are then defined to be

$$\widehat{\boldsymbol{\Sigma}}(0,1)^M = \sum_{j=1}^{L}\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^M = \sum_{j=1}^{L}\mathbf{P}_{-j}\text{diag}(\mathbf{P}_{-j}^{\text{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^M\mathbf{P}_{-j})\mathbf{P}_{-j}^{\text{T}}, \tag{3.16}$$

$$\widehat{\boldsymbol{\Sigma}}(0,1)^K = \sum_{j=1}^{L}\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^K = \sum_{j=1}^{L}\mathbf{P}_{-j}\text{diag}(\mathbf{P}_{-j}^{\text{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^K\mathbf{P}_{-j})\mathbf{P}_{-j}^{\text{T}}, \tag{3.17}$$

$$\widehat{\boldsymbol{\Sigma}}(0,1)^P = \sum_{j=1}^{L}\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^P = \sum_{j=1}^{L}\mathbf{P}_{-j}\text{diag}(\mathbf{P}_{-j}^{\text{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^P\mathbf{P}_{-j})\mathbf{P}_{-j}^{\text{T}}. \tag{3.18}$$

We also denote $\widehat{\boldsymbol{\Sigma}}(0,1)^{KP}$ and $\widehat{\boldsymbol{\Sigma}}(0,1)^{PP}$ the corresponding estimators using $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^{KP}$ and $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^{PP}$ respectively. An ideal estimator relative to $\widehat{\boldsymbol{\Sigma}}(0,1)$ is an estimator with $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^M$, $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^K$ (or $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^{KP}$) and $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^P$ (or $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^{PP}$)

replaced by the population counterpart $\mathbf{\Sigma}(\tau_{j-1}, \tau_j)$,

$$\mathbf{\Sigma}_{\text{Ideal}}(0,1) = \sum_{j=1}^{L} \mathbf{P}_{-j} \text{diag}(\mathbf{P}_{-j}^{\text{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{P}_{-j}) \mathbf{P}_{-j}^{\text{T}}. \tag{3.19}$$

As discussed in the paragraph after (3.12), if each partition is small, then say each $\mathbf{P}_{-j}$ is close to $\mathbf{P}$ which is the orthogonal matrix from the eigen-decomposition of $\sum_\ell \widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell)$, then the above becomes

$$\mathbf{\Sigma}_{\text{Ideal}}(0,1) \approx \mathbf{P} \text{diag}(\mathbf{P}^{\text{T}} \mathbf{\Sigma}(0,1) \mathbf{P}) \mathbf{P}^{\text{T}},$$

which resembles an ideal estimator for $\mathbf{\Sigma}(0,1)$ that utilizes all data information for constructing the eigenmatrix $\mathbf{P}$. Indeed such an ideal estimator is defined in Ledoit and Wolf (2012) and Lam (2016), and is treated as a benchmark for evaluating performances of different estimators. Note that if $\mathbf{P}$ is close to the eigenmatrix of $\mathbf{\Sigma}(0,1)$ (say, when $p/n \to 0$), then $\mathbf{\Sigma}_{\text{Ideal}}(0,1) \approx \mathbf{\Sigma}(0,1)$. As $p/n \to c > 0$ and $\mathbf{P}$ is getting further away from the true eigenmatrix of $\mathbf{\Sigma}(0,1)$, $\mathbf{\Sigma}_{\text{Ideal}}(0,1)$ in (3.19) is still close to $\mathbf{P}\text{diag}(\mathbf{P}^{\text{T}} \mathbf{\Sigma}(0,1) \mathbf{P}) \mathbf{P}^{\text{T}}$, the ideal rotation-equivariant estimator utilizing all information for the construction of $\mathbf{P}$.

In practice, this is the reason why we want each partition to be as small as possible, but with enough data points (we suggest at least over a hundred) such that the results in our theorems are reasonable. In all simulations and real data analysis in Section 3.6, the period $[0,1]$ represents a 5-day interval, and we use 1 day as the length of a partition (with typically hundreds of data points in each day), with very good results. From our experience, reducing the length of each partition (provided it still has enough data points, say over a hundred) usually improve the overall performance of the estimator. See Section 3.5 for more details on how to choose tuning parameters. In the following section, important assumptions in our paper are discussed before we present and explain the main theoretical results.

3.4 Asymptotic Theory

We introduce some notations first. Let the log-price series $\{\mathbf{X}_t\}$ be adapted to the filtered probability space $\left(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{0 \le t \le 1}, \mathbb{P}\right)$. We use $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ to denote the minimum and maximum eigenvalue of a square matrix respectively. Also, if no ambiguity arise, for $j = 1, \ldots, L$ we denote $v_s = v_s^j$ which is the sth all-refresh time

within partition j, and define

$$\mathcal{F}_{-j} = \mathcal{F}_{\tau_{j-1}} \cup \mathcal{F}/\mathcal{F}_{\tau_j}, \quad \mathcal{F}_s^j = \mathcal{F}_{v_s}/\mathcal{F}_{\tau_{j-1}},$$

with $\mathcal{F}_s^j = \phi$ for $s \leq 0$. They represent the σ-algebra outside of partition j, and that up to time v_s within partition j, respectively. We also define, for $j = 1, \ldots, L$,

$$\mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1}, \tau_j) = \mathbf{P}_{-j}\text{diag}(\mathbf{P}_{-j}^{\text{T}}\mathbf{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{P}_{-j})\mathbf{P}_{-j}^{\text{T}}, \tag{3.1}$$

so that $\mathbf{\Sigma}_{\text{Ideal}}(0, 1) = \sum_j \mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1}, \tau_j)$. We first introduce the assumptions for our theorems to hold, followed by some explanations. They are essentially the same as the main assumptions in Lam and Feng (2018). We present here for the sake of completeness and ease of reading for those readers who are interested in these assumptions.

In all assumptions below, we have

$$K = \begin{cases} K_m = N + m, & m = 1, \ldots, M, \text{ for the multi-scale method;} \\ 1, & \text{for the kernel method;} \\ 1, & \text{for the pre-averaging method.} \end{cases}$$

Assumptions on the drift $\boldsymbol{\mu}_t$:

(D1) The drift $\boldsymbol{\mu}_t$ has càdlàg components , such that for $s = K, K+1, \ldots, n(j)$,

$$\int_{v_{s-K}}^{v_s} \boldsymbol{\mu}_t dt = \mathbf{A}(v_{s-K}, v_s)\mathbf{Z}_{d,s}^j,$$

where $\mathbf{A}(v_{s-K}, v_s) \neq \mathbf{0}$ is a non-random $p \times p$ matrix, has $\left\|\mathbf{A}(v_{s-K}, v_s)\right\| = O(p_f^{1/2}K^{1/2}|v_s - v_{s-1}|)$ and can be asymmetric and singular, where $p_f = 1$ if there are no pervasive factors, and $p_f = p$ if there are pervasive factors. Also, $E(\mathbf{Z}_{d,s}^j|\mathcal{F}_{-j}) = \mathbf{0}$ and $\text{var}(\mathbf{Z}_{d,s}^j|\mathcal{F}_{-j}) = \mathbf{I}_p$ almost surely. The random vector $\mathbf{Z}_{d,s}^j \in \mathcal{F}_s^j$ has components conditionally independent of each other given $\mathcal{F}_{-j}$, with eighth moments exist. The drift $\boldsymbol{\mu}_t$ can also be non-random when $\mathbf{Z}_{d,s}^j = (1, 0, \ldots, 0)^T$ for all s.

(D2) Write $\mathbf{P}_{-j} = (\mathbf{p}_{1j}, \ldots, \mathbf{p}_{pj})$. We assume for each $i = 1, \ldots, p$, and $s = rK + q$ for $r = 1, \ldots, |S^j(K)|_K$ and $q = 0, 1, \ldots, K-1$, there exists $\rho_{d,K,q}^j \in \mathcal{F}_{-j}$ such

that $0 \leq \rho^j_{d,K,q} \leq \xi < 1$ with ξ a constant, and for $\ell = K+q, 2K+q, \ldots, rK+q$,

$$
\begin{aligned}
&E\big((\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{Z}^j_{d,\ell})^2 | \mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K}\big) \\
&= \rho^j_{d,K,q}(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{Z}^j_{d,\ell-K})^2 \\
&\qquad\qquad + (1-\rho^j_{d,K,q})\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{A}(v_{s-K}, v_s)^{\mathrm{T}}\mathbf{p}_{ij} + e^{ij}_{d,\ell-K},
\end{aligned}
$$

where we define $\mathbf{Z}^j_{d,\ell}\mathbf{Z}^{j\mathrm{T}}_{d,\ell} = \mathbf{I}_p$ and $e^{ij}_{d,\ell} = 0$ for $\ell \leq 0$. The process $\{e^{ij}_{d,\ell}\}$ with $e^{ij}_{d,\ell} \in \mathcal{F}^j_\ell$ has $E(e^{ij}_{d,\ell}|\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K}) = 0$ almost surely, and $e^{ij}_{d,\ell}|\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K} = O_P(\big\|\mathbf{A}(v_{s-K}, v_s)\big\|^2)$.

(D3) Let $\varphi(x) = e^{x^2} - 1$. We assume that for $\ell = 0, 1, \ldots, s$,

$$
\begin{aligned}
&E\bigg\{\varphi\bigg(\frac{|(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{Z}^j_{d,\ell})^2 - \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{A}(v_{s-K}, v_s)^{\mathrm{T}}\mathbf{p}_{ij}|}{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{Z}^j_{d,\ell-K})^2}\bigg)\Big|\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K}\bigg\} < \infty, \\
&E\bigg\{\varphi\bigg(\frac{|e^{ij}_{d,\ell}|}{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{Z}^j_{d,\ell-K})^2}\bigg)\Big|\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K}\bigg\} < \infty.
\end{aligned}
$$

Assumptions on the volatility $\boldsymbol{\sigma}_t$ and Brownian motion $\mathbf{W}_t$:

(V1) The volatility $\boldsymbol{\sigma}_t$ has càdlàg components, and the Brownian motion $\{\mathbf{W}_t\}$ can be correlated with $\{\boldsymbol{\mu}_t\}$ in general. Write

$$
\int_{v_{s-K}}^{v_s} \boldsymbol{\sigma}_t d\mathbf{W}_t = \boldsymbol{\Sigma}(v_{s-K}, v_s)^{1/2}\mathbf{Z}^j_{v,s},
$$

where $\boldsymbol{\Sigma}(v_{s-K}, v_s)$ is a symmetric positive definite $p \times p$ matrix which can be random, with

$$
\begin{aligned}
&\lambda_{\min}(\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)) \geq C(\tau_{j-1} - \tau_j)^{-1}, \\
&\lambda_{\max}(\boldsymbol{\Sigma}(v_{s-K}, v_s)) \asymp \big\|\mathbf{A}(v_{s-K}, v_s)\big\|^2/|v_s - v_{s-K}|,
\end{aligned}
$$

where $C > 0$ is a constant. The process $\{\boldsymbol{\sigma}_t\}$ is independent of all other processes.

Also, $E(\mathbf{Z}^j_{v,s}|\mathcal{F}_{-j}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{Z}^j_{v,s}|\mathcal{F}_{-j}) = \mathbf{I}_p$ almost surely. The random vector $\mathbf{Z}^j_{v,s} \in \mathcal{F}^j_s$ has components conditionally independent of each other given $\mathcal{F}_{-j}$, with eighth moments exist. The $\mathbf{Z}^j_{v,s}$'s are independent of each other for a fixed j.

(V2) Parallel to (D2), but expectations are taken conditional on $\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K} \cup \mathcal{F}^{\boldsymbol{\sigma}}_{v_s}$, where $\mathcal{F}^{\boldsymbol{\sigma}}_t$ is the σ-algebra generated by the process $\{\boldsymbol{\sigma}_t\}$ up to time t.

Also, $\rho^j_{d,K,q}$ is replaced by 0, $\mathbf{A}(v_{s-K}, v_s)$ by $\boldsymbol{\Sigma}(v_{s-K}, v_s)^{1/2}$, and $e^{ij}_{d,\ell}$ by 0.

(V3) Parallel to (D3), replacements the same as in (V2).

Assumptions on the microstructure noise $\boldsymbol{\epsilon}_t$:

(E1) Within the jth partition, $E(\boldsymbol{\epsilon}(s)\boldsymbol{\epsilon}(s)^{\mathrm{T}}|\mathcal{F}_{-j}) = \boldsymbol{\Sigma}^j_{\epsilon,s}$, which is random and independent of all other processes given $\mathcal{F}_{-j}$. Also, $E(\boldsymbol{\Sigma}^j_{\epsilon,s}) = \boldsymbol{\Sigma}^j_{\epsilon}$, and $\left\|\boldsymbol{\Sigma}^j_{\epsilon,s}\right\| \leq \lambda_\epsilon < \infty$ uniformly as $n, p \to \infty$ where λ_ϵ is a constant. The series $\{\boldsymbol{\Sigma}^j_{\epsilon,s}\}_s$ also satisfies a smoothness condition that $\left\|\boldsymbol{\Sigma}^j_{\epsilon,s+l} - \boldsymbol{\Sigma}^j_{\epsilon,s}\right\| \leq |l|/n$ for each $j = 1, \ldots, L$.

(E2) Within the jth partition, we can write $\boldsymbol{\epsilon}(s) = (\boldsymbol{\Sigma}^j_{\epsilon,s})^{1/2}\mathbf{Z}^j_{\epsilon,s}$, with $\mathbf{Z}^j_{\epsilon,s} \in \mathcal{F}^j_s$ having conditionally independent components given $\mathcal{F}_{-j}$. Also $E(\mathbf{Z}^j_{\epsilon,s}|\mathcal{F}_{-j}) = 0$ almost surely and eighth order moments exist for the components of $\mathbf{Z}^j_{\epsilon,s}$.

(E3) Let $\mathcal{F}^X_t$ be the σ-algebra generated by the log-price process up to time t, and $\mathcal{F}^\epsilon_t$ the one by the microstructure noise process up to time t, so that $\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}^X_s \otimes \mathcal{F}^\epsilon_s$. Then for s_1, s_2 time points within partition j, given $\mathcal{F}_{-j}$, we assume the φ-mixing coefficient between two σ-algebras satisfies

$$\varphi(\mathcal{F}^X_{s_1}, \mathcal{F}^\epsilon_{s_2}|\mathcal{F}_{-j}) = O(n^{-1}) = \varphi(\mathcal{F}^\epsilon_{s_2}, \mathcal{F}^X_{s_1}|\mathcal{F}_{-j}).$$

Also, for $s_2 > s_1$ time points within partition j, we assume

$$\varphi(\mathcal{F}^\epsilon_{s_1}, \mathcal{F}^\epsilon_{s_2}/\mathcal{F}^\epsilon_{s_1}|\mathcal{F}_{-j}) = O(n^{-1}) = \varphi(\mathcal{F}^\epsilon_{s_2}/\mathcal{F}^\epsilon_{s_1}, \mathcal{F}^\epsilon_{s_1}|\mathcal{F}_{-j}).$$

Other assumptions:

(A1) The observation times are independent of $\mathbf{X}(\cdot)$ and $\boldsymbol{\epsilon}(\cdot)$, and the partition boundaries τ_ℓ, $\ell = 0, 1, \ldots, L$, satisfy $0 < C_1 \leq \min_{\ell=1,\ldots,L} L(\tau_\ell - \tau_{\ell-1}) \leq \max_{\ell=1,\ldots,L} L(\tau_\ell - \tau_{\ell-1}) \leq C_2 < \infty$, where C_1, C_2 are generic constants. Also, the all-refresh times v_s, $s = 1, \ldots, nL$ satisfy $\max_{s=1,\ldots,nL} nL(v_s - v_{s-1}) \leq C_3$ for a generic constant $C_3 > 0$. Moreover, $\max_{\ell=1,\ldots,L} nL(\tau_\ell - v_{n(\ell)}) = o(1)$. The sample size in the jth partition has $n(j)/n \to 1$.

(A2) The pervasive factors, if any, persist within an interval $(v_{s-1}, v_s]$ for $s = 1, \ldots, nL$.

There are three more assumptions (A3), (A4) and (A5). They involve the drift and volatility in between the all-refresh and previous-tick times, and are in many ways parallel to assumptions (D1) to (D3) and (V1) to (V3). We present them in Section 3.7 to aid the flow of the paper. Assumptions (D1) to (D3) and (V1) to (V3) are mainly for the application of a version of Hoeffding's inequality on the sums of martingale differences in Theorem 2.2 of van de Geer (2002), and are also used in Lam and Feng (2018).

The matrix $\mathbf{A}(v_{s-K}, v_s)$ in assumptions (D1) to (D3) can be treated as a factor loading matrix in a factor model when $\boldsymbol{\mu}_t$ is random. The loading matrix $\mathbf{A}(v_{s-K}, v_s)$ is diagonal when the contribution of drift among all assets over v_{s-K} to v_s are conditionally independent given $\mathcal{F}_{-j}$. If there exists a factor structure with r factors where $r \ll p$, the loading matrix $\mathbf{A}(v_{s-K}, v_s)$ becomes a singular matrix whose rank is r. The first r singular values of $\mathbf{A}(v_{s-K}, v_s)$ are then of order $p^{1/2}K^{1/2}|v_s - v_{s-1}|$, with $K^{1/2}|v_s - v_{s-1}|$ accounting for the length of the time interval considered.

Also as introduced in Lam and Feng (2018), Assumption (D2) is about the serial drift dependence quantified by $\rho^j_{d,K,q} < 1$ given $\mathcal{F}_{-j}$. Assumption (D3) says that quadratic forms not too far in time apart can be very different but with sub-Gaussian-tailed probability. When $\boldsymbol{\mu}_t$ is non-random, the matrix $\mathbf{A}(v_{s-K}, v_s)$ can be set as zero except for a non-zero known vector in its first column. For more details please see the corresponding explanations in Lam and Feng (2018).

The assumptions for volatility from (V1) to (V3) have similar forms and meaning, and are parallel to the drift assumptions. One major difference from Assumption (V1) to (V3) in Lam and Feng (2018) is that we use the fact that the $\{\mathbf{Z}^j_{v,s}\}$'s are conditionally independent given $\mathcal{F}_{-j}$, which is in fact a consequence of independent increment in the Brownian motion $\{\mathbf{B}_t\}$. With this, Assumption (V2) certainly becomes

$$E\big((\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(v_{s-K}, v_s)^{1/2}\mathbf{Z}^j_{v,\ell})^2|\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K}\big) = \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(v_{s-K}, v_s)\mathbf{p}_{ij},$$

meaning that $\rho^j_{d,K,q}$ is replaced by 0, $e^{ij}_{d,\ell}$ by 0 and $\mathbf{A}(\cdot,\cdot)$ replaced by $\boldsymbol{\Sigma}(\cdot,\cdot)^{1/2}$. Also, note that Assumption (V1) actually ties the rate of the maximum eigenvalue of $\boldsymbol{\Sigma}(v_{s-K}, v_s)$ to $\big\|\mathbf{A}(v_{s-K}, v_s)\big\|$, meaning that if there are pervasive factors such that it spikes the singular value of $\mathbf{A}(\cdot,\cdot)$ by a factor of $p^{1/2}$, then the same factors also spikes the maximum eigenvalues of $\boldsymbol{\Sigma}(\cdot,\cdot)$. It makes sense since pervasive factors should have effects on both the drift and the volatility.

Assumption (E1) allows for time-varying covariance matrix for the microstructure noise. A smoothness condition for the time-varying covariance matrix is needed for

the bias-corrected pre-averaging method to successfully remove the residual covariance effects from the microstructure noise. Assumption (E3) particularly assumes a weak dependence between the log-price process and the microstructure noise process within partition j, as well as a weak serial dependence among the microstructure noise vectors, when $\mathcal{F}_{-j}$ is given. This assumption is inspired by Chen and Mykland (2017), where they assumed that given the entire information of the log-price process, the microstructure noise at different time points are independent. In our case, we are not given the entire picture of the log-price process, but not far from that either since with $\mathcal{F}_{-j}$ we are given $nL - n(j)$ data points from the total of nL. Then instead of assuming the microstructure noise vectors are independent, we assume that they are weakly dependent, and with n larger (i.e., more information at more time points is available outside partition j), the dependence is weaker.

The first part of Assumption (A1) is automatically satisfied if the boundary set $\{\tau_\ell\}_{0\le\ell\le L}$ is pre-set, for instance, to be the daily opening or closing time of the L days of data, or a quarter of it. See also Section 2.7 on a criterion in choosing these tuning parameters. Assumption (A2) means that the pervasive factors are either present between two all-refresh times, or they are absent.

Theorem 1. *Let Assumptions (D1) to (D3), (V1) to (V3), (E1) to (E3) and (A1) to (A5) hold. Then as $n, p \to \infty$ such that $p/n \to c \ge 0$,*

$$\max_{j=1,\dots,L} \left\| \widehat{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)^M \mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1},\tau_j)^{-1} - \mathbf{I}_p \right\| = \| \widehat{\mathbf{\Sigma}}(0,1)^M \mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1} - \mathbf{I}_p \|$$
$$= O_P(n^{-1/6} + p_f^{1/2} n^{-1/2}),$$

where $\|\cdot\|$ denotes the spectral norm of a matrix, and $p_f = 1$ if there are no factors, while $p_f = p$ if there are pervasive factors in the log-price processes.

The proof of the theorem is in Section 3.7. It is known that the rate of convergence for univariate realized volatility estimator in Zhang (2006) or the rate for multi-scale volatility matrix estimator under sparse assumption in Kim et al. (2016) is $n^{-1/4}$, the best achievable one. For our NER-MSRVM estimator, we have the slower rate $n^{-1/6}$ when there are no pervasive factors because we are not assuming sparsity of $\mathbf{\Sigma}(\tau_{j-1},\tau_j)$, although we still have the same rate as NERIVE in Lam and Feng (2018). In (3.4), the scales we use are $K_m = m + N$, $m = 1,\dots,M$. While Zhang (2006) and Kim et al. (2016) use $N \asymp n^{1/2}$ with $M \asymp n^{1/2}$, we use $N \asymp n^{2/3}$ and $M \asymp n^{1/2}$, so that $K_m \asymp n^{2/3}$, the same magnitude as the scale used in Zhang (2011) and Lam and Feng (2018) in their two-scale estimator. We find that this larger scale is needed to remove the adverse effects of large p without sparsity while removing the bias from microstructure noise.

When there are pervasive factors, $p_f = p$, and the rate tell us that we need $p = o(n)$ in order to still have guaranteed convergence as $n, p \to \infty$. What happens is that some drift terms can somehow dominates the volatility terms when there are pervasive factors which spike up the maximum singular value of $\mathbf{A}(\cdot,\cdot)$ and $\mathbf{\Sigma}(\cdot,\cdot)^{1/2}$ simultaneously by an order of $p^{1/2}$. Ultimately, when $p_f = p$ which has the same order as $n^{2/3}$, then we still have $n^{-1/6}$ as the rate of convergence, which coincides with the results for NERIVE in Lam and Feng (2018). If the factors are not pervasive, then $p_f = p^{1-\delta}$ for some $0 < \delta < 1$ which represents the strength of the factors (Lam and Yao, 2012, Lam et al., 2011). Clearly we can still have $p/n \to c > 0$ in such a case, and can retain the rate $n^{-1/6}$ if $\delta \geq 1/3$.

In practice, a larger M means we are using more scales, which can improve the estimation in general. This aligns with our simulation results which show that NER-MSRVM performs better than NERIVE in Lam and Feng (2018) which uses only two scales.

Theorem 2. *With the same conditions as in Theorem 1, as $n, p \to \infty$ such that $p/n \to c \geq 0$, we have*

$$\max_{j=1,\ldots,L} \left\| \widehat{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)^K \mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1}, \tau_j)^{-1} - \mathbf{I}_p \right\| = \| \widehat{\mathbf{\Sigma}}(0,1)^K \mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1} - \mathbf{I}_p \|$$
$$= O_P(p_f^{1/2} n^{-1/4}),$$

with $H \asymp n^{1/2}$ and $J \asymp p_f^{-1/2} n^{1/4}$, where $p_f = 1$ if there are no factors, and $p_f = p$ if there are pervasive factors. For the positive semi-definite version, we have

$$\max_{j=1,\ldots,L} \left\| \widehat{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)^{KP} \mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1}, \tau_j)^{-1} - \mathbf{I}_p \right\| = \| \widehat{\mathbf{\Sigma}}(0,1)^{KP} \mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1} - \mathbf{I}_p \|$$
$$= O_P(p_f^{2/5} n^{-1/5}),$$

with $H \asymp p_f^{-1/5} n^{3/5}$ and $J \asymp p_f^{-2/5} n^{1/5}$ respectively.

The proof of this theorem is in Section 3.7. Kim et al. (2016) introduces KRVM which has a faster rate of convergence at $n^{-1/4}$, and KRPVM, which has rate only at $n^{-1/5}$ but is guaranteed to be positive semi-definite. The above theorem shows that NER-KRVM is approaching $\mathbf{\Sigma}_{\text{Ideal}}(0,1)$ in spectral norm at a rate of $n^{-1/4}$ when there are no factors, and hence it is positive definite in probability. The same goes for NER-KRPVM, but the rate is slower at $n^{-1/5}$. When there are pervasive factors in the log-price processes, all rates of convergence to the ideal estimator are slower. As explained after Theorem 1, this happens since some terms related to the drift can

dominate when there are pervasive factors. This theorem suggests that for our regularized estimator for the kernel method, the bias-corrected version NER-KRVM is always better than NER-KRPVM since they are both positive definite in probability anyway, while the former converges faster to the ideal estimator $\mathbf{\Sigma}_{\text{Ideal}}(0,1)$. This is true as long as $p = o(n^{1/2})$, when $p^{2/5}n^{-1/5}$ is going to 0 slower than $p^{1/2}n^{-1/4}$.

Theorem 3. *With the same conditions as in Theorem 1, and $n, p \to \infty$ such that $p/n \to c \geq 0$,*

$$\max_{j=1,\ldots,L} \left\| \widehat{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)^P \mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1},\tau_j)^{-1} - \mathbf{I}_p \right\| = \| \widehat{\mathbf{\Sigma}}(0,1)^P \mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1} - \mathbf{I}_p \|$$
$$= O_P(p_f^{3/8} n^{-1/4}),$$

with $Q \asymp p_f^{-1/4} n^{1/2}$, where $p_f = 1$ if there are no factors, and $p_f = p$ if there are pervasive factors. We also have

$$\max_{j=1,\ldots,L} \left\| \widehat{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)^{PP} \mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1},\tau_j)^{-1} - \mathbf{I}_p \right\| = \| \widehat{\mathbf{\Sigma}}(0,1)^{PP} \mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1} - \mathbf{I}_p \|$$
$$= O_P(p_f^{2/5} n^{-1/5}),$$

with $Q \asymp p_f^{-1/5} n^{3/5}$.

The proof of this theorem can be found in Section 3.7. Same as kernel estimators, Kim et al. (2016) also gives two versions of pre-averaging estimators: the bias-corrected version PRVM, which is not guaranteed to be positive semi-definite but is converging at a rate of $n^{-1/4}$, and the positive semi-definite version PRPVM, which has a slower rate of convergence at $n^{-1/5}$. Our results show that the regularized estimators are all positive definite in probability, and when there are no factors, the rate of convergence to the ideal estimator $\mathbf{\Sigma}_{\text{Ideal}}(0,1)$ for NER-PRVM is $n^{-1/4}$. For NER-PRPVM, it is $n^{-1/5}$ under $p/n \to c \geq 0$. These rates are the same as those for the non-regularized estimators with finite p. Same as Theorem 1 and 2, some terms related to the drift can dominate as $p \to \infty$, and hence ultimately all rates of convergence are slower. However, NER-PRVM still has a faster rate of convergence to the ideal estimator at $p^{3/8}n^{-1/4}$ when compared to that for NER-PRPVM at $p^{2/5}n^{-1/5}$. Note also that $p = o(n^{2/3})$ is needed for NER-PPVM while $p = o(n^{1/2})$ is needed for NER-PRPVM for guaranteed convergence when there are pervasive factors.

Comparing the rate $p_f^{3/8}n^{-1/4}$ for NER-PRVM to $p_f^{1/2}n^{-1/4}$ for NER-KRVM, it is clear that when there are pervasive factors, the pre-averaging estimator can converge

a bit faster to the ideal estimator. It also allows $p = o(n^{2/3})$ compared to only $p = o(n^{1/2})$ allowed for NER-KRVM. The practical performance for NER-PRVM is also better than NER-KRVM in many scenarios as demonstrated in Section 3.6.1. The simulation results also highlight that NER-PRVM is better than NER-PRPVM in general. All our regularized estimators constructed during simulations and real data analysis are also positive definite, showing that the convergence is in fact pretty quick. In the following section, we also show that our regularized estimators are adaptive to jumps removal.

3.4.1 Jumps Remove

In Section 3.2.1, a continuous-diffusion price model 3.1 is used as our framework. Although microstructure noise has been included in our method, jump should also be considered. Compared with most of existing methods developed for either noisy data from a continuous-diffusion price model or data from a jump diffusion price model without noise, Fan and Wang (2007) proposed methods to cope with both jumps in the price and market microstructure noise in the observed data. The idea is to remove jumps from the data first and apply noise-resistant methods to do estimation. We use a same idea but under a high dimensional scenario.

Considering jumps, we suppose a the underlying log-price process as follow

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\sigma}_t d\mathbf{W}_t + d\mathbf{J}_t, \quad t \in [0, 1], \tag{3.2}$$

where $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$ are same as the pure diffusion model (3.1), and $\mathbf{J}_t = (J_t^{(1)}, \ldots, J_t^{(p)})^{\mathrm{T}}$ is a p-dimensional right-continuous pure jump process. For jth asset,

$$J_t^{(j)} = \sum_{\ell=1}^{N_t^{(j)}} B_\ell^{(j)}, \quad t \in [0, 1],$$

where each count process $N_t^{(j)}$ can be correlated with each other. The same holds true for each jump size $B_\ell^{(j)}$.

We assume that the number of jumps in each $X_t^{(j)}$ over the time period we consider is finite. Then the quadratic covariation over $[0, 1]$ for log process is denoted as

$$QV = \int_0^1 \boldsymbol{\sigma}_t \boldsymbol{\sigma}_t^{\mathrm{T}} dt + \sum_{0 \le t \le 1} \Delta \mathbf{J}_t \Delta \mathbf{J}_t^{\mathrm{T}}, \tag{3.3}$$

where $\Delta \mathbf{J}_t = \mathbf{J}_t - \mathbf{J}_{t-}$. When two log price processes have at least one cojumps simutaneously, the corresponding element in $\Delta \mathbf{J}_t \Delta \mathbf{J}_t^{\mathrm{T}}$ will be non-zero. This can happen when a shock news affects some stocks at same time.

Same as Lam and Feng (2018), we use the wavelet method introduced in Fan and Wang (2007) to remove the jumps in the log price process and then build the volatility matrix estimator as 3.16, 3.17 and 3.18 by jump-removed data. To apply this wavelet method successfully, more assumptions are need:

(W1) The wavelets used in jump estimation are differentiable.

(W2) For the jump part of $X_t^{(j)}$ in $[0,1]$ for $j = 1, \ldots, p$, its jump locations $\eta_\ell^{(j)}$ and jump sizes $B_\ell^{(j)}$ satisfy

$$N_1^{(j)} < \infty,\ \eta_1^{(j)} < \ldots < \eta_\ell^{(j)} < \ldots,\ 0 < |B_\ell^{(j)}| < \infty \text{ almost surely.}$$

(W3) The number of stocks involved in a cojump is finite.

Assumptions (W1) and (W2) are technical assumptions adapted from Fan and Wang (2007). Assumption (W2) means that we are dealing with finite number of jumps for each log-price process, and the sizes of the jumps are bounded from 0 almost surely. If Assumption (W3) is not satisfied, then the rate of convergence of $\widehat{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)$ in Theorem 1 and 3 using the jumps-removed data will be dependent on how many stocks is involved in a cojump in general. Our assumptions allow the jump process to be dependent on the drift, volatility and the microstructure noise process in general.

Theorem 4. *Let all the assumptions in Theorem 1 hold, as well as (W1) to (W3) for the jump-diffusion model (3.2). Using the jumps-removed all-refresh log-price data* $\mathbf{Y}^*(s) = \mathbf{Y}(s) - \widehat{\mathbf{J}}_{v_s}$, $s = 1, \ldots, nL$ *in constructing the integrated covariance matrix estimator in (3.13), (3.14) or (3.15), the same conclusions in Theorem 1, Theorem 3 and their semi-positive definiteness hold. Moreover, we have*

$$\Big\| \sum_{0 \le t \le 1} (\Delta \mathbf{J}_t \Delta \mathbf{J}_t^T - \Delta \widehat{\mathbf{J}}_t \Delta \widehat{\mathbf{J}}_t^T) \Big\| = O_P(n^{-1/4}).$$

3.5 Practical Implementation

There are two parameters that can be tuned for potentially better performance, namely the partition $(\tau_{j-1}, \tau_j]$ of the period $[0,1]$ (thus also determining L itself which represents the number of partitions), the scale parameter M and N used in

the MSCV in (3.4), the window length H and jittering length J in the KRCV in (3.6) and the averaging range K in the PRCV in (3.9).

As for the number of partitions, different number of partitions corresponds to different length of each individual partition. Therefore, same as Lam and Feng (2018), we propose a criterion as follow:

$$g(\boldsymbol{\tau}) = \left\| \sum_{j=1}^{L} \left(\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j) - \widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j) \right) \right\|_F^2, \tag{3.1}$$

where $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^M$, $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^K$ and $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^P$ defined in (3.4), (3.6) and (3.9). Similarly, $\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^M$, $\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^K$ and $\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^P$ are shown in (3.13), (3.14) and (3.15) respectively. Bickel and Levina (2008) proposed this function firstly to choose the banding number in large covariance matrix estimator with banding structure assumption. Lam and Feng (2018) suggested to divide the time interval into equal length partitions, checking that each one has enough data points and then choose L by minimizing the criterion (3.1) above.

With respect to other tuning parameters, the scale parameter M and N, the window length H, jittering length J and the averaging range K, we already find the theoretical magnitude of all of them related to sample size n showing in 3.4, 3.6 and 3.9. To final decide the value of these tuning parameters, constant part for all of them are set as 1, as in our simulation and real data analysis show that the results are not very sensitive to the choice of the constant.

3.6 Empirical Results

3.6.1 Simulations

To generate simulated data for high frequency problem, the prices and the asynchronous transaction times should be generated separately. We apply Heston-like multivariate factor model with stochastic volatilities used in Lam and Feng (2018) as follow:

$$dX_t^{(i)} = \mu^{(i)} dt + \sqrt{1 - (\rho^{(i)})^2}\sigma_t^{(i)} dB_t^{(i)} + \rho^{(i)}\sigma_t^{(i)} dW_t + C\nu^{(i)} dZ_t, \quad i = 1, \ldots, 100, \tag{3.2}$$

where $\{W_t\}$, $\{Z_t\}$ and the $\{B_t^{(i)}\}$'s are independent standard Brownian motions. The processes $\{W_t\}$ and $\{Z_t\}$ imitate factors in the market. The constant $C = 1_{\{\text{model 2}\}}$

is 0 for the first model we consider. We set $\rho^{(i)} = -0.7C$, so that it is 0 in the first model, and hence there are no factors. For the second model, $C = 1$, so that it contains two factors. The spot volatility $\sigma_t^{(i)} = \sqrt{\varrho_t^{(i)}}$ follows the Cox-Ingersoll-Ross (CIR) process

$$d\varrho_t^{(i)} = \kappa^{(i)}(\theta^{(i)} - \varrho_t^{(i)})dt + \xi^{(i)}dU_t^{(i)},$$

where the $\{U_t^{(i)}\}$'s are independent standard Brownian motions. Other parameters of $X_t^{(i)}$ are set at $(\mu^{(i)}, \kappa^{(i)}, \xi^{(i)}, \theta^{(i)}) = (0.03x_1^{(i)}, 1.1x_2^{(i)}, 0.5x_3^{(i)}, 0.25x_4^{(i)})$ and $\nu^{(i)} = \sqrt{\theta^{(i)}}$, where the $x_j^{(i)}$'s are independent uniform random variables on the interval $[0.7, 1.3]$. The initial value of each log-price $X_0^{(i)}$ is set randomly on the interval $[0.5, 1.5]$ and the starting spot volatility $\varrho_0^{(i)}$ on the interval $[0.5\theta^{(i)}, 1.5\theta^{(i)}]$.

To consider the microstructure noise, as shown in (3.2), the simulated observed log-price is $X_t^{o(i)} = X_t^{(i)} + \varepsilon_t^{(i)}$, where $X_t^{(i)}$ represents the latent log-price, and the microstructure noise has $\varepsilon_t^{(i)} \overset{iid}{\sim} N(0, 0.0005^2)$. For the transaction times, we generate 100 different Poisson processes with intensities $\lambda_1, \ldots, \lambda_{100}$ respectively. To exam the finite sample performance for our proposed methods, we assume that one day is 23400 seconds, λ_i is set to be $0.01i \times 23400$, where $i = 1, \ldots, 100$.

We have three estimators NER-MSRVM in (3.16), NER-KRVM in (3.17) and NER-PRVM in (3.18). As shown in Theorem 1, the largest scale M becomes $n^{2/3}$ from the original $n^{1/2}$. To exam the necessity of this change in high dimensional setting, a modified NER-MSRVM, as NER-mMSRVM, with a smaller $M \asymp n^{1/2}$ is also considered as a competitor. In Theorem 1, the convergence rate for NER-MSRVM is same as the corresponding rate shown in Lam and Feng (2018) for NER-TSRVM. In order to see the difference in the finite sample example, the NER-TSRVM introduced in Lam and Feng (2018) is also included into our simulation.

The nonparametric eigenvalue regularization introduced in (3.3.4) is applied to reduce the negative effect from high dimensional setting. Therefore, the corresponding estimators without the nonparametric eigenvalue regularization are also included here and denoted as TSRVM, MSRVM, mMSRVM, KRVM and PRVM for NER-TSRVM, NER-MSRVM, NER-mMSRVM, NER-KRVM and NER-PRVM respectively.

Besides the methods above, Dai et al. (2017) proposed a POET method based on pre-averaging data to capture the factor structure and handle the microstructure noise. We apply this method on our simulated data and denote it as PR-POET. Furthermore, a pure POET method without pre-averaging step is also included in our simulation.

Section 3.4.1 introduces the jump-diffusion process and the jump removal process for our proposed methods. Our simulation generate two set of data. The first

scenario is for no jump process, which follows Heston-like multivariate factor model as (3.2). The other is about jump included process obtained by adding one or three jumps on each price and each time interval in the first scenario process. The jump location uniformly distributed on the time interval, which is 5-days interval in our simulation. Then the jump size follows iid normal distribution with mean 0 and standard deviation 1/30. This jump adding procedure is proposed in Fan and Wang (2007) and the standard deviation 1/30 is used in their paper.

As for the measurement in our simulation, we use Frobenius error defined by $\mathrm{tr}(\widehat{\mathbf{\Sigma}}(0,1)-\mathbf{\Sigma}(0,1))^2$. The integrated covariance matrix $\mathbf{\Sigma}(0,1)$ is evaluated using the simulated latent log-prices at the finest grid (1 per second). We divide the 100 trading days into disjoint 5-day intervals, and calculate the Frobenius error for different estimators over each 5-day interval.

Figure 3.1 to Figure 3.4 are simulation results for no factor model where $C = 0$. In Figure 3.1, firstly, the proposed nonparametric eigenvalue regularization performs well in terms of removing the incluence from high dimension, as all regularized estimators, NER-TSRVM, NER-MSRVM, NER-KRVM and NER-PRVM outperform the conrresponding non-regularized ones. Secondly, our proposed methods, NER-MSRVM, NER-KRVM and NER-PRVM have smaller Frobenius errors in general, and especially NER-PRVM performs best. Thirdly, there is no significant influence when we include jumps and jump removal method proposed in Fan and Wang (2007).

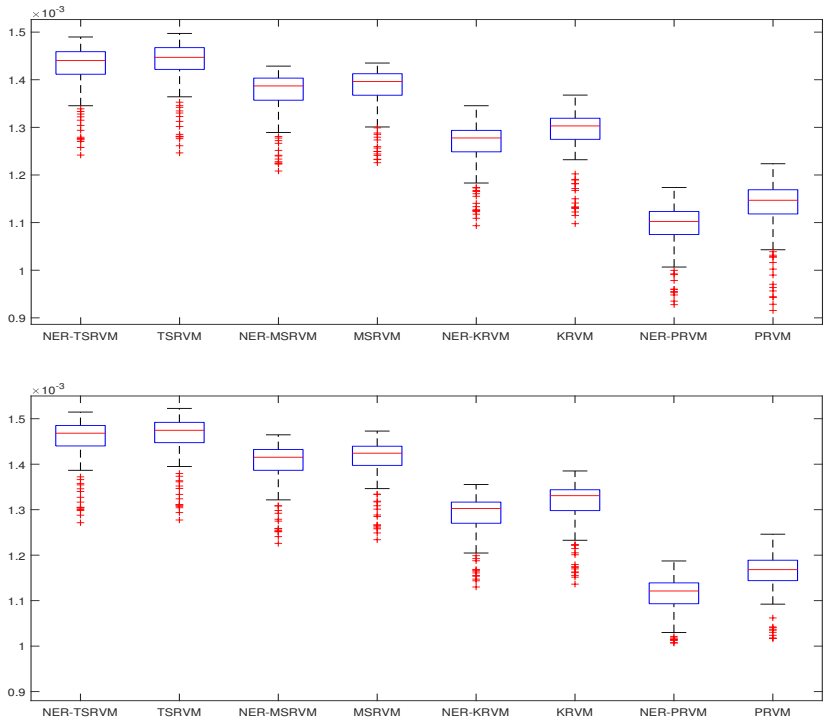


Figure 3.1: Boxplots of Frobenius errors of NER-TSRVM, TSRVM, NER-MSRVM, MSRVM, NER-KRVM, KRVM, NER-PRVM and PRVM for $C = 0$. The upper plot is for no jump scenario, while the bottom one is for jumps model (sd=1/30) result.

Both NER-MSVRM and NER-mMSVRM are considered in our simulation. Figure 3.2 reflects that, unlike the magnitude of parameter N in Zhang (2006), a larger magnitude 2/3 is needed in high dimensional volatility matrix estimation. It is easy to find that NER-MSRVM is significantly better than NER-mMSRVM no matter whether there is no jump or jumps in the model. As the simulation with jumps has same results, we omit it here to save space.

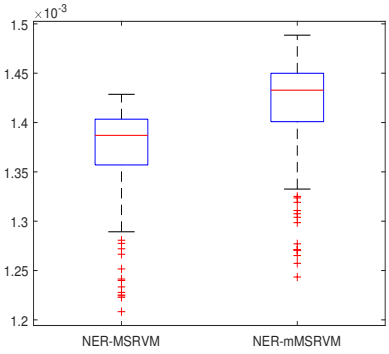


Figure 3.2: Boxplots of Frobenius errors of MSRVM and mMSRVM with nonparametric regularization for $C = 0$.

In Kim et al. (2016), besided the bias corrected estimators KRVM and PRVM, KRPVM and PRPVM are also discussed as they are positive semi-definite but have

slower convergent rate in theory. We also consider both bias corrected estimators and positive definite estimators as shown in Section 3.3.4. When there is no jump in the model, NER-KRVM and NER-PRVM outperform NER-KRPVM and NER-PRPVM respectively, which matches our theory. However, if the jumps are included, after jump removal procedure proposed in Fan and Wang (2007), the results are similar as Figure 3.3.

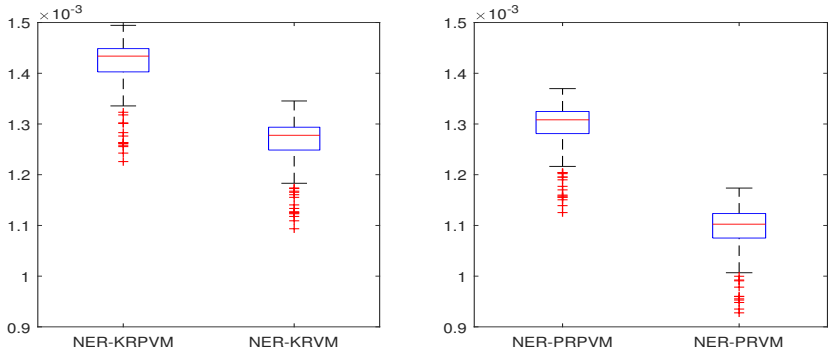


Figure 3.3: Boxplots of Frobenius errors of positive semi-definite estimators (NER-pKRVM and NER-pPRVM) and bias-corrected estimators (NER-KRVM and NER-PRVM) for $C = 0$.

To compare with POET and PR-POET introduced in Dai et al. (2017), by Figure 3.4, we can find that NER-PRVM are much better than PR-POET and POET. As both NER-PRVM and PR-POET apply pre-averaging method, the difference between NER-PRVM and PR-POET show us that the nonparametric eigenvalue regularizaion proposed in this paper can fix bias caused by high dimension better.

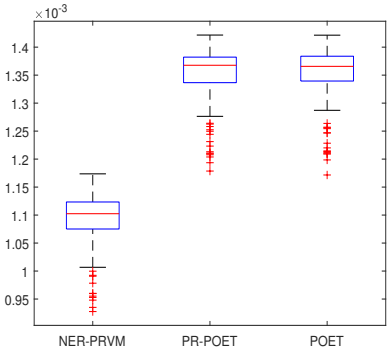


Figure 3.4: Boxplots of Frobenius errors of NER-PRVM, PR-POET and POET for $C = 0$.

Figure 3.5 to 3.8 provide the results for factor model where $C = 1$. All results are same as the conclusion made in Figure 3.1 to 3.4.

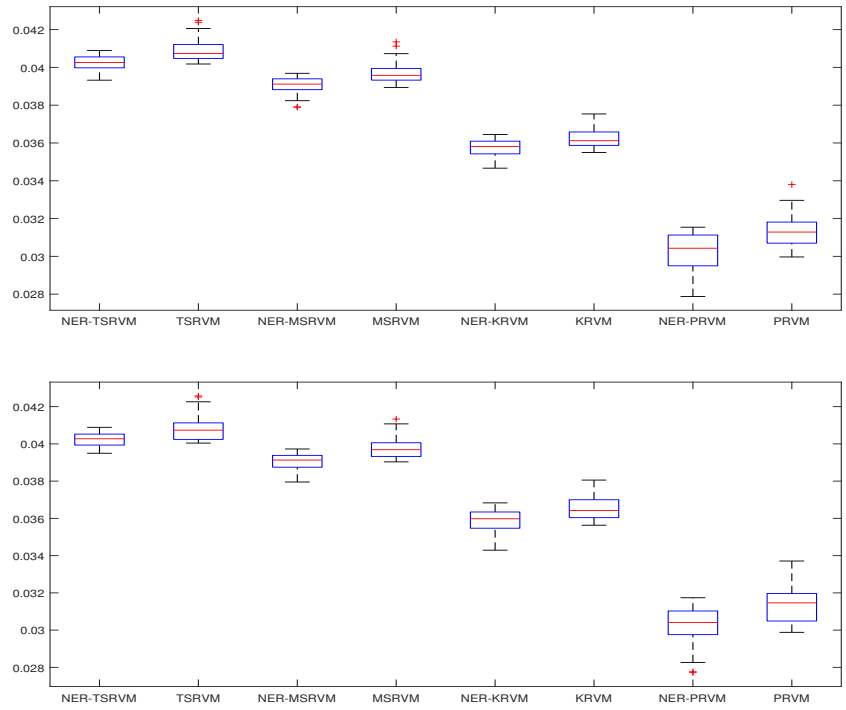


Figure 3.5: Boxplots of Frobenius errors of TSRVM, MSRVM, KRVM, PRVM and their nonparametric regularization estimators for $C = 1$. The upper plot is for no jump scenario, while the bottom one is for jumps model (sd=1/30) result.

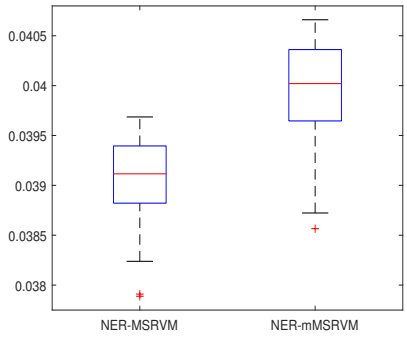


Figure 3.6: Boxplots of Frobenius errors of MSRVM and mMSRVM (scale is 1/2) with nonparametric regularization for $C = 1$.

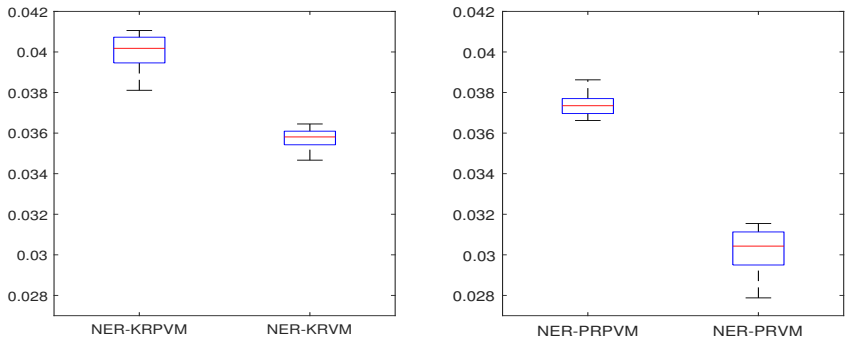


Figure 3.7: Boxplots of Frobenius errors of positive semi-definite estimators (NER-pKRVM and NER-pPRVM) and bias-corrected estimators (NER-KRVM and NER-PRVM) for $C = 1$.

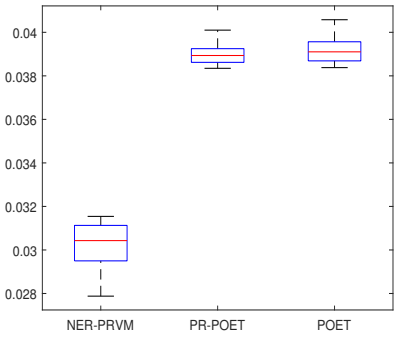


Figure 3.8: Boxplots of Frobenius errors of NER-PRVM, PR-POET and POET for $C = 1$.

As we know that one of challenges in the problem we want to solve are microstructure noise, Figure 3.9 indicates the results for the proposed methods and the competitors when the microstructure noise variance is changing from small to large value. For both factor model or the model without factor structure, when a small microstructure noise is included, NER-PRVM and NER-KRVM perform better than NER-MSRVM and NER-TSRVM. Then, with the increase of noise, NER-MSRVM can still obtain a good results as it dose when a small noise is used, while the Frobenius errors for NER-PRVM and NER-TSRVM diverge hugely.

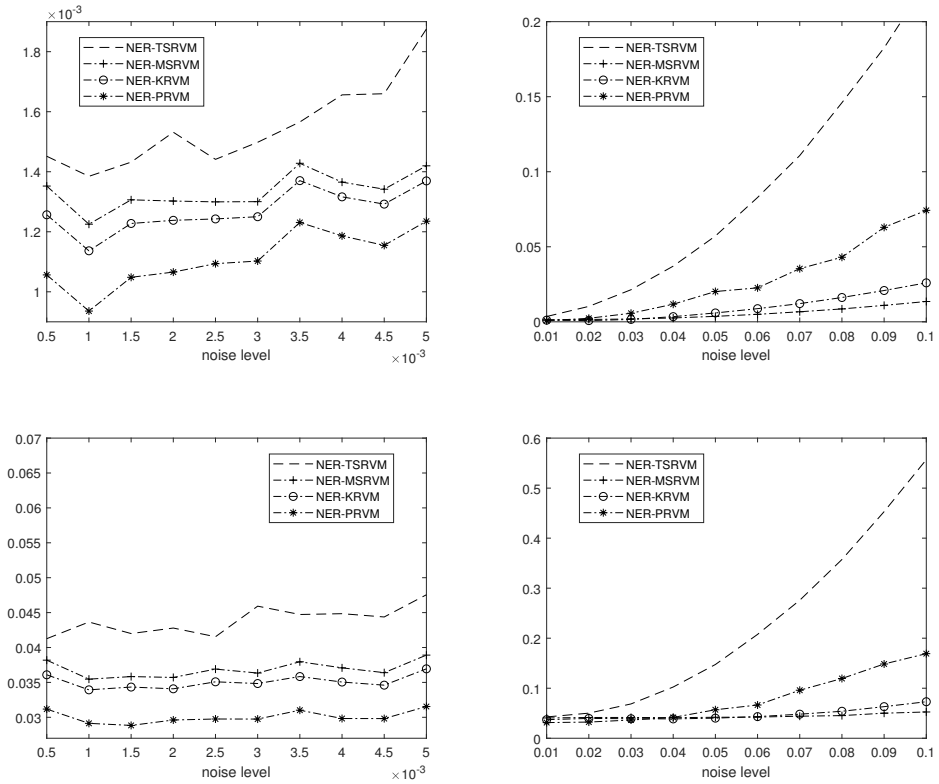


Figure 3.9: Simulation results about microstructure noise effect by Frobenius errors for model 3.2 without factors ($C = 0$) and with factors ($C = 1$) from the upper to the bottom but no jumps.

Another question we concern is about high dimensionality. Figure 3.10 provides the simulation results when p increase from 10 to 100. With increasing p, as the larger matrix makes question more challenging, all methods included become worse obviously in Figure 3.10. But we can find that NER-PRVM performs better.

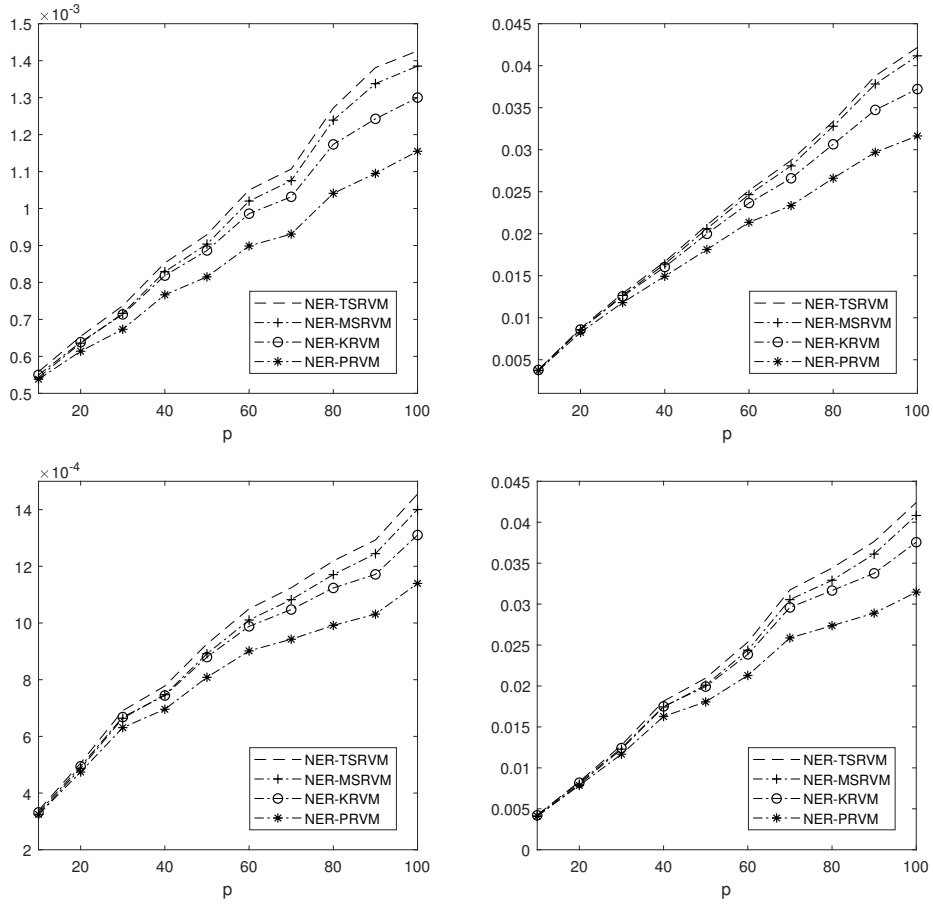


Figure 3.10: Simulation results about dimension p effect by Frobenius errors for model 3.2 without factors ($C = 0$) and with factors ($C = 1$) from left to right with no jump and with jumps from upper to the bottom.

3.6.2 Real Data

3.6.2.1 Minimum variance portfolio allocation

We first present a Theorem on minimum variance portfolio using our integrated volatility matrix estimators. Define $\mathbf{1}_p$ a column vector of p ones, the theoretical minimum variance portfolio has weights defined by

$$\mathbf{w}_{\text{theo}} = \frac{\boldsymbol{\Sigma}(0,1)^{-1}\mathbf{1}_p}{\mathbf{1}_p^{\mathrm{T}}\boldsymbol{\Sigma}(0,1)^{-1}\mathbf{1}_p}.$$

An estimated minimum variance portfolio weight vector is then

$$\hat{\mathbf{w}} = \frac{\widehat{\boldsymbol{\Sigma}}(0,1)^{-1}\mathbf{1}_p}{\mathbf{1}_p^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}(0,1)^{-1}\mathbf{1}_p}.$$

We also define the maximum exposure of a portfolio $\mathbf{w}$ as $\|\mathbf{w}\|_{\max} = \max_i |w_i|$.

Theorem 5. *Let all assumptions in Theorem 3 hold. If there are no pervasive factors with $p/n \to c > 0$, or there are pervasive factors but $p^{3/2}/n \to c > 0$, the maximum exposures of $\mathbf{w}_{\text{theo}}$ and $\hat{\mathbf{w}}$, with $\widehat{\mathbf{\Sigma}}(0,1)$ being equal to $\widehat{\mathbf{\Sigma}}(0,1)^M$, $\widehat{\mathbf{\Sigma}}(0,1)^K$, $\widehat{\mathbf{\Sigma}}(0,1)^{KP}$, $\widehat{\mathbf{\Sigma}}(0,1)^P$ or $\widehat{\mathbf{\Sigma}}(0,1)^{PP}$, will satisfy in probability*

$$p^{1/2}\|\hat{\mathbf{w}}\|_{\max},\ p^{1/2}\|\mathbf{w}_{\text{theo}}\|_{\max} \leq \frac{\max_{1\leq j\leq L}\lambda_{\max}(\mathbf{\Sigma}(\tau_{j-1},\tau_j))}{\min_{1\leq j\leq L}\lambda_{\min}(\mathbf{\Sigma}(\tau_{j-1},\tau_j))}.$$

If there are no pervasive factors and $p/n \to c > 0$, the actual risk of $\hat{\mathbf{w}}$ (the same as above) and $\mathbf{w}_{\text{theo}}$, with an actual risk of $\mathbf{w}$ defined by $R^{1/2}(\mathbf{w}) = (\mathbf{w}^T\mathbf{\Sigma}(0,1)\mathbf{w})^{1/2}$, satisfy in probability

$$p^{1/2}R^{1/2}(\hat{\mathbf{w}}) \leq \frac{\max_{1\leq j\leq L}\lambda_{\max}(\mathbf{\Sigma}(\tau_{j-1},\tau_j))}{\min_{1\leq j\leq L}\lambda_{\min}(\mathbf{\Sigma}(\tau_{j-1},\tau_j))}\cdot\lambda^{1/2}_{\max}(\mathbf{\Sigma}(0,1)),$$
$$p^{1/2}R^{1/2}(\mathbf{w}_{\text{theo}}) \leq \lambda^{1/2}_{\max}(\mathbf{\Sigma}(0,1)).$$

If there are pervasive factors and $p^{3/2}/n \to c > 0$, then $R(\hat{\mathbf{w}}) = O_P(\lambda(\mathbf{\Sigma}(0,1))) = O_P(p)$, and the bound for $R(\mathbf{w}_{\text{theo}})$ remains the same as above.

If Assumptions (W1) to (W3) hold also under the jump-diffusion model (3.2), then the same conclusions as above hold, as long as we are using the jumps removal procedure described in Section 3.4.1.

Theorem 5 is exactly the same as Theorem 5 in Lam and Feng (2018), and the proof is also the same and hence interested readers are encouraged to read the proof of Theorem 5 there. The maximum exposure bound is important since it is clear that the theoretical minimum variance portfolio also satisfies this bound. In practice, unless we are using other methods and building towards a concentrated portfolio, as far as minimum portfolio is concerned, we do not want to invest in a single asset too much in any single period of time, especially when the theoretical portfolio is not doing so.

3.6.2.2 NYSE data analysis

In this study, we choose the stocks based on two lists, the "Fifty Most Active Stocks on NYSE, Round Lots (mils. of shares), 2013" and "Fifty Most Active Stocks by Dollar Volume on NYSE ($ in mils.), 2013", from the New York Stock Exchange Data official website `http://www.nyxdata.com/`. There are 26 stocks appearing in

both of the lists above, and 74 stocks in either of them. We downloaded all the trading transactions of these 74 stocks in Year 2013 from the Wharton Research Data Services (WRDS, `https://wrds-web.wharton.upenn.edu/`). We omit the stock Sprint Corporation due to missing price data. We first clean all the data by the R-package "highfrequency", which follows the high frequency data cleaning steps presented in Barndorff-Nielsen et al. (2009). We conduct our portfolio allocation study on two portfolios, one with the $p = 26$ stocks appearing in both lists, and the other with $p = 73$ stocks appearing in either of the lists.

The quantities to be compared for different portfolios are as follows. For daily rebalancing with a k-day training window ($k = 1$ or 5), we calculate the annualized portfolio return and annualized out-of-sample standard deviation, given respectively by

$$\hat{\mu} = 250 \times \frac{1}{250-k} \sum_{i=k+1}^{250} \mathbf{w}^{\mathrm{T}} \mathbf{r}_i, \quad \hat{\sigma} = \Big(250 \times \frac{1}{250-k} \sum_{i=k+1}^{250} (\mathbf{w}^{\mathrm{T}} \mathbf{r}_i - \frac{\hat{\mu}}{250})^2\Big)^{1/2}.$$

The out-of-sample standard deviation is a good indicator of how much risk is associated with a portfolio Demiguel and Nogales (2009), and is our main quantity for performance comparisons, whereas portfolio return is of secondary importance. We also calculate the Sharpe ratio $\hat{\mu}/\hat{\sigma}$. The average maximum exposure and the maximum of the maximum exposure over the whole investment horizon are two important measures for comparisons too. Since this is a simulation experiment, we can calculate the actual risk of a portfolio $\mathbf{w}$, $R^{1/2}(\mathbf{w}) = (\mathbf{w}^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{w})^{1/2}$, over a trading day. We compare the averaged actual risks of different methods over the whole investment horizon. Finally we compare the error norm compared to $\mathbf{w}_{\mathrm{theo}}$, defined as Norm $= \big\|\mathbf{w} - \mathbf{w}_{\mathrm{theo}}\big\|$, and also the portfolio turnover for different methods.

Before applying the proposed methods, we remove the jumps as shown in Section 3.4.1. It is clear to see that our nonparametric eigenvalue threshold methods perform well to reduce the side effect from high dimensionality, especially also better than POET method. In terms of averaged maximum absolute weight, maximum of maximum absolute weight, annualized portfolio return and Sharpe ratio, kernel method and pre-averaging method are better than two-scale and multi-scale estimators in general.

5-day Methods	Out-of-Sample SD (%)	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Return(%)	Sharpe Ratio
		p=26			
NER-TSRVM	4.7	$20_{(6)}$	44	17.1	3.63
TSRVM	5.7	$41_{(14)}$	94	16.3	2.85
NER-MSRVM	4.6	$20_{(6)}$	42	17.8	3.86
MSRVM	5.7	$42_{(14)}$	98	17.4	3.05
NER-mMSRVM	4.9	$21_{(7)}$	49	17.7	3.61
mMSRVM	5.7	$42_{(16)}$	108	17.4	4.39
NER-KRPVM	4.6	$19_{(5)}$	39	20.2	4.39
KRPVM	5.3	$27_{(7)}$	92	16.2	3.06
NER-KRVM	4.6	$17_{(6)}$	36	21.6	4.69
KRVM	5.2	$26_{(7)}$	89	15.3	2.94
NER-PRPVM	4.6	$18_{(4)}$	33	20.2	4.39
PRPVM	5.4	$24_{(6)}$	86	17.3	3.20
NER-PRVM	4.5	$17_{(4)}$	31	22.0	4.89
PRVM	5.3	$25_{(7)}$	82	14.3	2.70
POET	7.0	$33_{(37)}$	75	16.2	2.31
PR-POET	4.7	$26_{(6)}$	50	16.8	3.57
		p=72			
NER-TSRVM	4.0	$11_{(3)}$	22	14.9	3.72
TSRVM	141.7	$238_{(508)}$	4708	−456.5	−3.22
NER-MSRVM	4.0	$11_{(4)}$	23	14.4	3.6
MSRVM	137.5	$260_{(514)}$	4215	−474.5	−3.45
NER-mMSRVM	4.2	$12_{(4)}$	24	14.3	3.40
mMSRVM	157.8	$251_{(507)}$	4917	−460.9	−2.92
NER-KRPVM	4.0	$10_{(3)}$	21	14.9	3.72
KRPVM	137.5	$236_{(469)}$	4352	−413.5	−3.01
NER-KRVM	3.9	$11_{(5)}$	21	14.8	3.79
KRVM	140.6	$229_{(431)}$	4623	−398.7	−2.83
NER-PRPVM	3.9	$10_{(2)}$	22	15.9	4.08
PRPVM	125.2	$224_{(453)}$	4145	−416.7	−3.32
NER-PRVM	3.8	$9_{(2)}$	21	16.7	4.39
PRVM	128.1	$216_{(448)}$	4007	−398.5	−3.11
POET	62.7	$188_{(232)}$	3116	−235.4	−3.75
PR-POET	4.1	$16_{(6)}$	33	14.5	3.53

Table 3.1: Empirical results (jumps removed) for the 26 and 72 most actively traded stocks in NYSE: annualized out-of-sample standard deviation, averaged maximum absolute weight, maximum of maximum absolute weight, annualized portfolio return and Sharpe ratio.

3.7 Proof of Theorems

Before presenting any proofs, we present the last set of assumptions which are required for Theorem 1,2,3 and 4 to hold. We first need to decompose $\mathbf{X}_{v_s} - \mathbf{X}(s)$. Consider the previous-tick time $t_s^i \in (v_{s-1}, v_s]$ for the ith asset, which should satisfy

$$v_{s-1} < t_s^{i_1} \leq t_s^{i_2} \leq \cdots \leq t_s^{i_p} = v_s,$$

where $\{i_1, \ldots, i_p\}$ is some permutation of $1, \ldots, p$. Letting b_s (assumed $o(p)$) denote the number of tides, we can write the above as

$$v_{s-1} < t_s^{j_1} < t_s^{j_2} < \cdots < t_s^{j_{p-b_s}} = v_s,$$

where $j_1, \ldots, j_{p-b_s} \in \{1, \ldots, p\}$.

Then we can write, for $s = 1, \ldots, nL$,

$$\mathbf{X}_{v_s} - \mathbf{X}(s) = \sum_{m=1}^{p-b_s-1} \mathbf{D}_m^s \mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{Z}_{d,s}^j(m+1) + \sum_{m=1}^{p-b_s-1} \mathbf{D}_m^s \mathbf{\Sigma}(t_s^{j_m}, t_s^{j_{m+1}})^{1/2} \mathbf{Z}_{v,s}^j(m+1), \tag{3.1}$$

where $\mathbf{D}_m^s$ is a diagonal matrix with either 0 or 1 as elements. The jth diagonal element is 1 if the jth asset is already traded at time $t_s^{j_m}$, and 0 otherwise. The matrices $\mathbf{A}(\cdot, \cdot)$ and $\mathbf{\Sigma}(\cdot, \cdot)$ are as defined in Assumption (D1) and (V1) respectively.

(A3) If the drift $\boldsymbol{\mu}_t$ is random, the components of $\mathbf{Z}_{d,s}^j(m+1)$, $\mathbf{Z}_{v,s}^j(m+1) \in \mathcal{F}^j_{t_s^{j_{m+1}}}$ are conditionally independent given $\mathcal{F}_{-j}$, $E(\mathbf{Z}_{d,s}^j(m+1)|\mathcal{F}_{-j}) = \mathbf{0} = E(\mathbf{Z}_{v,s}^j(m+1)|\mathcal{F}_{-j})$, $\mathrm{var}(\mathbf{Z}_{d,s}^j(m+1)|\mathcal{F}_{-j}) = \mathbf{I}_p = \mathrm{var}(\mathbf{Z}_{v,s}^j(m+1)|\mathcal{F}_{-j})$ almost surely. Eighth moments exist for their components as well. If the drift $\boldsymbol{\mu}_t$ is non-random, then $\mathbf{Z}_{d,s}^j(m+1) = (1, 0, \ldots, 0)^T$.

(A4) (Only for random drift and volatility). Using notations in Assumption (D2), we assume that for some $c_{d,j,s} \in \mathcal{F}_{-j} \cup \mathcal{F}_s^j$ greater than 0, and for $\ell = 1, \ldots, m$,

$$\begin{aligned} &E\Big(\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{D}_m^s \mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{Z}_{d,s}^j(\ell+1) \Big| \mathcal{F}_{-j} \cup \mathcal{F}^j_{t_s^{j_\ell}}\Big) \\ &= \left(1 - \frac{c_{d,j,s}}{(p - b_s - 1)^\alpha}\right) \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{D}_m^s \mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{Z}_{d,s}^j(\ell) + e_{d,s}^{ij}(\ell), \end{aligned}$$

where $0 \leq \alpha \leq 1/2$, and we define $\mathbf{Z}_{d,s}^j(\ell) \mathbf{Z}_{d,s}^j(\ell)^{\mathrm{T}} = \mathbf{I}_p$ and $e_{d,s}^{ij}(\ell) = 0$ for $\ell \leq 0$. The process $\{e_{d,s}^{ij}(\ell)\}$ with $e_{d,s}^{ij}(\ell) \in \mathcal{F}^j_{t_s^{j_\ell}}$ has $E(e_{d,s}^{ij}(\ell)|\mathcal{F}_{-j} \cup \mathcal{F}^j_{t_s^{j_{\ell-1}}}) = 0$

almost surely, and $e^{ij}_{d,s}(\ell)|\mathcal{F}_{-j} \cup \mathcal{F}^j_{t_s^{j_{\ell-1}}} = O_P(\|\mathbf{A}(t_s^{j_{\ell-1}}, t_s^{j_\ell})\|) = O_P(p_f^{1/2} \cdot (p - b_s - 1)^{-1} n^{-1} L^{-1})$.

We also have $E\Big(\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{D}_m^s \boldsymbol{\Sigma}(t_s^{j_m}, t_s^{j_{m+1}})^{1/2} \mathbf{Z}^j_{v,s}(\ell+1) \Big| \mathcal{F}_{-j} \cup \mathcal{F}^j_{t_s^{j_\ell}} \cup \mathcal{F}^{\boldsymbol{\sigma}}_{v_s}\Big) = 0$, since $\{\mathbf{Z}^j_{v,s}(\ell)\}_{s,\ell}$ for $s = 1, \ldots, nL$, $\ell = 1, \ldots, p - b_s - 1$ is a double array of independent random vectors.

(A5) (Only for random drift and volatility). Let $\varphi(x) = e^{x^2} - 1$. We assume that for $\ell = 1, \ldots, m$,

$$E\left\{\varphi\left(\frac{|\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{D}_m^s \mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{Z}^j_{d,s}(\ell)|}{|\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{D}_m^s \mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{Z}^j_{d,s}(\ell-1)|}\right) \Big| \mathcal{F}_{-j} \cup \mathcal{F}^j_{t_s^{j_{\ell-1}}}\right\} < \infty,$$
$$E\left\{\varphi\left(\frac{|e^{ij}_{d,s}(\ell)|}{|\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{D}_m^s \mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{Z}^j_{d,s}(\ell-1)|}\right) \Big| \mathcal{F}_{-j} \cup \mathcal{F}^j_{t_s^{j_{\ell-1}}}\right\} < \infty.$$

The assumption for the volatility runs parallel to the above, with the expectations now conditional on $\mathcal{F}_{-j} \cup \mathcal{F}^j_{t_s^{j_{\ell-1}}} \cup \mathcal{F}^{\boldsymbol{\sigma}}_{v_s}$, $\boldsymbol{\Sigma}(\cdot,\cdot)^{1/2}$ replaces $\mathbf{A}(\cdot,\cdot)$, $\mathbf{Z}^j_{v,s}(\cdot)$ replaces $\mathbf{Z}^j_{d,s}(\cdot)$ and 0 replaces $e^{ij}_{d,s}(\cdot)$.

Assumptions (A3), (A4) and (A5) are parallel to (D1), (D2) and (D3) respectively. The major difference is that the coefficients $\rho^j_{d,K,q} < 1$ is now replaced by coefficients that are going to 1 as $n, p \to \infty$, which are controlled by an exponent $\alpha > 0$. A larger α means that the correlations among variables between tick-by-tick trading times are higher. We assume this since the time length between ticks is usually very small. Note that if the drift is non-random, we only need Assumption (A3) that $\mathbf{Z}^j_{d,s}(m+1) = (1, 0, \ldots, 0)^T$, which is just a matter of notation rather than an assumption.

If we assume the jump-diffusion model (3.2) for the log-price process $\mathbf{X}_t$, and all estimators are constructed using the jumps-removed data

$$\widetilde{\mathbf{Y}}_t = \mathbf{Y}_t - \widehat{\mathbf{J}}_t = (\mathbf{X}_t - \widehat{\mathbf{J}}_t) + \boldsymbol{\epsilon}_t = \widetilde{\mathbf{X}}_t + \boldsymbol{\epsilon}_t, \tag{3.2}$$

where $\{\widehat{\mathbf{J}}_t\}$ is the estimated jump process using the wavelet method in Fan and Wang (2007), then $\{\widetilde{\mathbf{X}}_t\}$ represents the jumps-removed log-price process. For $j = 1, \ldots, L$ and $v_s = v_s^j$ for $s = 0, \ldots, n(j)$, we then have

$$\widetilde{\mathbf{Y}}(s) = \widetilde{\mathbf{X}}(s) + \boldsymbol{\epsilon}(s) = \widetilde{\mathbf{X}}_{v_s} + \mathbf{E}(s),$$

where we define

$$\mathbf{E}(s) = \boldsymbol{\epsilon}(s) + \widetilde{\mathbf{X}}(s) - \widetilde{\mathbf{X}}_{v_s} = \boldsymbol{\epsilon}(s) + (\mathbf{X}(s) - \widehat{\mathbf{J}}(s)) - (\mathbf{X}_{v_s} - \widehat{\mathbf{J}}_{v_s}).$$

Replacing $\mathbf{X}(s)$ and $\mathbf{X}_{v_s}$ by $\mathbf{X}(s) + \mathbf{J}(s)$ and $\mathbf{X}_{v_s} + \mathbf{J}_{v_s}$ respectively where $\mathbf{X}_t$ is the vector of pure jump log-price process, we then have

$$\mathbf{E}(s) = \boldsymbol{\epsilon}(s) + \widetilde{\mathbf{X}}(s) - \widetilde{\mathbf{X}}_{v_s} = \boldsymbol{\epsilon}(s) + (\mathbf{X}(s) - \mathbf{X}_{v_s}) + (e(\mathbf{J}(s)) - e(\mathbf{J}_{v_s})),$$

where $e(\mathbf{J}_t) = \mathbf{J}_t - \hat{\mathbf{J}}_t$. Hereafter, we denote $\mathbf{E}(s)$ by the above expansion, and $\tilde{\mathbf{X}}_t = \mathbf{X}_t + e(\mathbf{J}_t)$.

Before we present the proofs of our theorems, we introduce five important lemmas first. The first two lemmas come from Lam and Feng (2018) and are used for proving Theorem 1, 2 and 3. The remaining three lemmas are developed for proving Theorem 2 and 3.

For any integer $m \geq 1$, define

$$\begin{aligned}
[\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(m)} &= \frac{1}{m} \sum_{s, s+m \in S^j(m)} (\widetilde{\mathbf{X}}_{v_{s+m}} - \widetilde{\mathbf{X}}_{v_s})(\widetilde{\mathbf{X}}_{v_{s+m}} - \widetilde{\mathbf{X}}_{v_s})^{\mathrm{T}}, \\
[\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(m)} &= \frac{1}{m} \sum_{s, s+m \in S^j(m)} (\widetilde{\mathbf{X}}_{v_{s+m}} - \widetilde{\mathbf{X}}_{v_s})(\mathbf{E}(s+m) - \mathbf{E}(s))^{\mathrm{T}}, \\
[\mathbf{E}, \mathbf{E}^{\mathrm{T}}]_j^{(m)} &= \frac{1}{m} \sum_{s, s+m \in S^j(m)} (\mathbf{E}(s+m) - \mathbf{E}(s))(\mathbf{E}(s+m) - \mathbf{E}(s))^{\mathrm{T}}.
\end{aligned} \tag{3.3}$$

Lemma 3. *Let all the assumptions in Theorem 4 hold. Then with $p/n \to c \geq 0$, we have*

$$\max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \frac{\mathbf{p}_{ij}^T [\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^T]_j^{(K_m)} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^T \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} - 1 \right| = O_P(K_m^{1/2} n^{-1/2} + p_f^{1/2} n^{-1/2}).$$

The proof for Lemma 3 can be found in the intermediate results in the proof of Lemma 1 in Lam and Feng (2018). We do not repeat their calculations here.

Lemma 4. *Let all the assumptions in Theorem 4 hold. Then with $p/n \to c \geq 0$, we have*

$$\max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \max_{s=1,\ldots,n(j)} \left| \frac{\mathbf{p}_{ij}^T (\mathbf{X}_{v_s} - \mathbf{X}(s))}{(\mathbf{p}_{ij}^T \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}} \right| = O_P(n^{-1/2}).$$

This Lemma is in fact the same as Lemma 2 in Lam and Feng (2018) where $\alpha = 1/6$ there. The different result stems from Assumption (V2) where we treat the $\mathbf{Z}^j_{v,s}(m)$ as an independent sequence of random vectors in both the indices m and s. We omit the proof here, since it is the same proof as Lemma 2 of Lam and Feng (2018), but with $\alpha = 0$ on the volatility side of the proof.

In all the proofs below, if no ambiguity arises, we use n in a summation instead of $n(j)$ within a partition j to simplify notations since we assumed $n \asymp n(j)$ for each j anyway.

Lemma 5. *Let all the assumptions in Theorem 4 hold. Then with $p/n \to c \geq 0$, we have for a bandwidth Q defined in (3.8),*

$$\max_{\substack{i=1,\ldots,p,\ j=1,\ldots,L \\ l=1,\ldots,Q-1,\ l'=1,\ldots,Q-1 \\ l\neq l'}} \left| \sum_{s=1}^{n-Q+1} \frac{\mathbf{p}_{ij}^T(\widetilde{\mathbf{X}}_{v_{s+l}} - \widetilde{\mathbf{X}}_{v_{s+l-1}})(\widetilde{\mathbf{X}}_{v_{s+l'}} - \widetilde{\mathbf{X}}_{v_{s+l'-1}})^T\mathbf{p}_{ij}}{\mathbf{p}_{ij}^T\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} \right| = O_P(p_f^{1/2}n^{-1/2}).$$

Proof of Lemma 5. We use the decomposition

$$\sum_{s=1}^{n-Q+1} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}_{v_{s+l}} - \widetilde{\mathbf{X}}_{v_{s+l-1}})(\widetilde{\mathbf{X}}_{v_{s+l'}} - \widetilde{\mathbf{X}}_{v_{s+l'-1}})^T\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = \frac{P^{(1)}_{l,l'} + 2P^{(2)}_{l,l'} + P^{(3)}_{l,l'}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

where

$$\begin{aligned}
P^{(1)}_{l,l'} &= \sum_{s=1}^{n-Q+1} (\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s+l-1},v_{s+l})\mathbf{Z}^j_{d,s+l} + \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_{s+l-1},v_{s+l})^{1/2}\mathbf{Z}^j_{v,s+l}) \\
&\qquad \cdot (\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s+l'-1},v_{s+l'})\mathbf{Z}^j_{d,s+l'} + \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_{s+l'-1},v_{s+l'})^{1/2}\mathbf{Z}^j_{v,s+l'})^{\mathrm{T}}, \\
P^{(2)}_{l,l'} &= \sum_{s=1}^{n-Q+1} (\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s+l-1},v_{s+l})\mathbf{Z}^j_{d,s+l} + \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_{s+l-1},v_{s+l})^{1/2}\mathbf{Z}^j_{v,s+l})(e(\mathbf{J}_{v_{s+l'}}) - e(\mathbf{J}_{v_{s+l'-1}}))^{\mathrm{T}}\mathbf{p}_{ij}, \\
P^{(3)}_{l,l'} &= \sum_{s=1}^{n-Q+1} \mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}_{v_{s+l}}) - e(\mathbf{J}_{v_{s+l-1}}))(e(\mathbf{J}_{v_{s+l'}}) - e(\mathbf{J}_{v_{s+l'-1}}))^{\mathrm{T}}\mathbf{p}_{ij}.
\end{aligned}$$

Let us consider $P_{l,l'}^{(1)}$ first. We know that

$$\begin{aligned}
P_{l,l'}^{(1)} =& \sum_{s=1}^{n-Q+1} (\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s+l-1}, v_{s+l}) \mathbf{Z}_{d,s+l}^{j} \mathbf{Z}_{d,s+l'}^{j\mathrm{T}} \mathbf{A}(v_{s+l'-1}, v_{s+l'})^{\mathrm{T}} \mathbf{p}_{ij}) \\
&+ \sum_{s=1}^{n-Q+1} (\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s+l-1}, v_{s+l}) \mathbf{Z}_{d,s+l}^{j} \mathbf{Z}_{v,s+l'}^{j\mathrm{T}} \mathbf{\Sigma}(v_{s+l'-1}, v_{s+l'})^{\mathrm{T}/2} \mathbf{p}_{ij}) \\
&+ \sum_{s=1}^{n-Q+1} (\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(v_{s+l-1}, v_{s+l})^{1/2} \mathbf{Z}_{v,s+l}^{j} \mathbf{Z}_{d,s+l'}^{j\mathrm{T}} \mathbf{A}(v_{s+l'-1}, v_{s+l'})^{\mathrm{T}} \mathbf{p}_{ij}) \\
&+ \sum_{s=1}^{n-Q+1} (\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(v_{s+l-1}, v_{s+l})^{1/2} \mathbf{Z}_{v,s+l}^{j} \mathbf{Z}_{v,s+l'}^{j\mathrm{T}} \mathbf{\Sigma}(v_{s+l'-1}, v_{s+l'})^{\mathrm{T}/2} \mathbf{p}_{ij}) \\
=& P_{l,l'}^{(1,1)} + P_{l,l'}^{(1,2)} + P_{l,l'}^{(1,3)} + P_{l,l'}^{(1,4)}.
\end{aligned}$$

Consider $P_{l,l'}^{(1,1)}$. The proof is the same with non-random drift where then $p_f = p$ since $\mathbf{A}(\cdot,\cdot)$ contains just one non-zero column of vector. So here we only focus on proof with random drift. We have

$$\begin{aligned}
E_j|P_{l,l'}^{l,l'}| \leq& \sum_{s=1}^{n-Q+1} E_j^{1/2} (\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s+l-1}, v_{s+l}) \mathbf{Z}_{d,s+l}^{j})^2 E_j^{1/2} (\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s+l'-1}, v_{s+l'}) \mathbf{Z}_{d,s+l'}^{j})^2 \\
=& O(\left\|\mathbf{A}(v_{s+l-1}, v_{s+l})\right\| \left\|\mathbf{A}(v_{s+l'-1}, v_{s+l'})\right\|) = O(p_f n^{-1}).
\end{aligned}$$

Also, using Lemma 2.7 of Bai and Silverstein (1998),

$$\begin{aligned}
E_j|P_{l,l'}^{l,l'}|^2 \leq& \sum_{s=1}^{n-Q+1} E_j^{1/2} (\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s+l-1}, v_{s+l}) \mathbf{Z}_{d,s+l}^{j})^4 E_j^{1/2} (\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s+l'-1}, v_{s+l'}) \mathbf{Z}_{d,s+l'}^{j})^4 \\
&+ \sum_{s_1 \neq s_2} E_j((\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s_1+l-1}, v_{s_1+l}) \mathbf{Z}_{d,s_1+l}^{j} \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s_2+l-1}, v_{s_2+l}) \mathbf{Z}_{d,s_2+l}^{j}) \\
&\quad \cdot (\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s_1+l'-1}, v_{s_1+l'}) \mathbf{Z}_{d,s_1+l'}^{j} \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s_2+l'-1}, v_{s_2+l'}) \mathbf{Z}_{d,s_2+l'}^{j})) \\
=& O(n^2 \cdot \left\|\mathbf{A}(v_{s+l-1}, v_{s+l})\right\|^4) = O(p_f^2 n^{-2}).
\end{aligned}$$

Hence we have

$$P_{l,l'}^{(1,1)} = O_P(p_f n^{-1}).$$

Same technique shows that

$$P_{l,l'}^{(1,2)}, P_{l,l'}^{(1,3)} = O_P(p_f n^{-1/2}).$$

Now consider $P^{(1,4)}_{l,l'}$. By Assumption (V1), for $\ell' < \ell$, defining

$$R_{\ell,\ell'} = \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(v_{\ell-1}, v_\ell)^{1/2} \mathbf{Z}^j_{v,\ell} \mathbf{Z}^{j\mathrm{T}}_{v,\ell'} \mathbf{\Sigma}(v_{\ell'-1}, v_\ell)^{\mathrm{T}/2} \mathbf{p}_{ij},$$

we can use the independence of the $\mathbf{Z}^j_{v,s}$'s to see that $E(R_{\ell,\ell'}|\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-1}) = 0$. Hence using the Burkholder's inequality, we have

$$E_j(P^{(1,4)}_{l,l'})^2 \leq C \sum_{s=1}^{n-Q+1} E_j(R^2_{s+l,s+l'}).$$

where C is a generic constant. Using Lemma 2.7 of Bai and Silverstein (1998),

$$\begin{aligned} E(R^2_{\ell,\ell'}) &\leq 2E\left(R^2_{\ell,\ell} + R^2_{\ell',\ell'}\right), \quad \text{with} \\ E_j(R^2_{\ell,\ell}) &= O(\left\|\mathbf{\Sigma}(v_{\ell-1}, v_\ell)\right\|^2) = O(p_f^2 n^{-2}). \end{aligned}$$

Hence we can conclude that $P^{(1,4)}_{l,l'} = O_P(p_f n^{-1/2})$, and hence

$$\begin{aligned} P^{(1)}_{l,l'} / \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij} &= (P^{(1,1)}_{l,l'} + P^{(1,2)}_{l,l'} + P^{(1,3)}_{l,l'} + P^{(1,4)}_{l,l'}) / \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij} \\ &= O_P(p_f^{1/2} n^{-1/2}). \end{aligned}$$

Using the jumps removal rate in Fan and Wang (2007), and Assumption (W2) and (W3) that there are only finite number of jumps for each stock and cojumps at each time point, we have

$$P^{(3)}_{l,l'} / \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij} = O_P(n^{-1/2}).$$

For $P^{(2)}_{l,l'}$, similar to the treatment above,

$$P^{(2)}_{l,l'} = O_P(p_f^{1/2} n^{-5/4} + p_f^{1/2} n^{-3/4}),$$

which is dominated by $P^{(1)}_{l,l'}$. Hence we have

$$\max_{\substack{i=1,\ldots,p,\ j=1,\ldots,L \\ l=1,\ldots,Q-1,\ l'=1,\ldots,Q-1 \\ l \neq l'}} \left| \sum_{s=1}^{n-Q+1} \frac{\mathbf{p}_{ij}^{\mathrm{T}} (\widetilde{\mathbf{X}}_{v_{s+l}} - \widetilde{\mathbf{X}}_{v_{s+l-1}})(\widetilde{\mathbf{X}}_{v_{s+l'}} - \widetilde{\mathbf{X}}_{v_{s+l'-1}})^T \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} \right| = O_P(p_f^{1/2} n^{-1/2}). \quad \square$$

Lemma 6. *Let all the assumptions in Theorem 4 hold. Then with $p/n \to c \geq 0$,*

$$\max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \sum_{s=1}^{n} \frac{\mathbf{p}_{ij}^{T}(\mathbf{E}(s)\mathbf{E}(s)^{T})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{T}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} \right| = O_P(n).$$

Proof of Lemma 6. By the definition of $\mathbf{E}(s)$, we have

$$\sum_{s=1}^{n} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{E}(s)\mathbf{E}(s)^{\mathrm{T}})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = \sum_{\ell=1}^{3} E_\ell + 2\sum_{\ell=4}^{6} E_\ell, \text{ where}$$

$$E_1 = \sum_{s=1}^{n} \frac{(\mathbf{p}_{ij}^{\mathrm{T}}(\boldsymbol{\epsilon}(s)))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

$$E_2 = \sum_{s=1}^{n} \frac{(\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s) - \mathbf{X}_{v_s}))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

$$E_3 = \sum_{s=1}^{n} \frac{(\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}(s)) - e(\mathbf{J}_{v_s}))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

$$E_4 = \sum_{s=1}^{n} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\boldsymbol{\epsilon}(s))(\mathbf{X}(s) - \mathbf{X}_{v_s})^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

$$E_5 = \sum_{s=1}^{n} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\boldsymbol{\epsilon}(s))(e(\mathbf{J}(s)) - e(\mathbf{J}_{v_s}))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

$$E_6 = \sum_{s=1}^{n} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s) - \mathbf{X}_{v_s})(e(\mathbf{J}(s)) - e(\mathbf{J}_{v_s}))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}.$$

By Lemma 4, and Assumption (W2) and (W3) together with the rate of jumps removal in Fan and Wang (2007), we have

$$E_2 = O_P(1),\ E_3 = O_P(n^{-1/2}).$$

For E_1, consider $E_1 = E_{1,1} + E_{1,2}$, where

$$E_{1,1} = \frac{\sum_{s=1}^{n}((\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s))^2 - \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij})}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

$$E_{1,2} = \frac{\sum_{s=1}^{n}\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}. \tag{3.4}$$

For $E_{1,2}$, Assumption (E1) implies that

$$E_{1,2} = O_P(n).$$

For $E_{1,1}$, let $g_s^{ij} = (\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s))^2 - \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij}$. Using Lemma 2.7 of Bai and Silverstein (1998) under Assumption (E1) to (E3), we have

$$\begin{aligned}
&E(E_{1,1}^2|\mathcal{F}_{-j} \cup \{\boldsymbol{\Sigma}_{\epsilon,u}, u \in [0,1]\}) \\
&= (\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{-2}\Big\{\sum_{s=1}^{n} E((g_s^{ij})^2|\mathcal{F}_{-j} \cup \{\boldsymbol{\Sigma}_{\epsilon,u}, u \in [0,1]\}) \\
&\quad + \sum_{s_1 \neq s_2} E(g_{s_1}^{ij} g_{s_2}^{ij}|\mathcal{F}_{-j} \cup \{\boldsymbol{\Sigma}_{\epsilon,u}, u \in [0,1]\})\Big\} \\
&= O(n \cdot 1 + n^2 \cdot n^{-1} \cdot 1) = O(n).
\end{aligned}$$

Hence $E_{1,1} = O_P(n^{1/2})$, and so

$$E_1 = O_P(n).$$

Finally, a simple C-r inequality shows that either E_1, E_2 or E_3 dominates the order of E_4 to E_6, and hence

$$\sum_{s=1}^{n} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{E}(s)\mathbf{E}(s)^{\mathrm{T}})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = O_P(n). \quad \square$$

Lemma 7. *Let all the assumptions in Theorem 4 hold. Then with $p/n \to c \geq 0$, for $0 < l < n$,*

$$\sum_{l'=0}^{l-1}\sum_{s=1}^{n-l} \frac{\mathbf{p}_{ij}^{T}\mathbf{E}(s+l)\mathbf{E}(s+l')^{T}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{T}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = O_P(l^{1/2}n^{1/2}).$$

Proof of Lemma 7. We can decompose $\sum_{l'=0}^{l-1}\sum_{s=1}^{n-l} \frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}(s+l)\mathbf{E}(s+l')^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = \sum_{j=1}^{9} L_j$,

where

$$L_1 = \sum_{s=1}^{n-l} \frac{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s+l) \sum_{l'=0}^{l-1} \boldsymbol{\epsilon}(s+l')^{\mathrm{T}} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}},$$

$$L_2 = \sum_{s=1}^{n-l} \frac{\mathbf{p}_{ij}^{\mathrm{T}} (\mathbf{X}(s+l) - \mathbf{X}_{v_{s+l}}) \sum_{l'=0}^{l-1} (\mathbf{X}(s+l') - \mathbf{X}_{v_{s+l'}})^{\mathrm{T}} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}},$$

$$L_3 = \sum_{s=1}^{n-l} \frac{\mathbf{p}_{ij}^{\mathrm{T}} (e(\mathbf{J}(s+l)) - e(\mathbf{J}_{v_{s+l}})) \sum_{l'=0}^{l-1} (e(\mathbf{J}(s+l')) - e(\mathbf{J}_{v_{s+l'}}))^{\mathrm{T}} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}},$$

$$L_4 = \sum_{s=1}^{n-l} \frac{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s+l) \sum_{l'=0}^{l-1} (\mathbf{X}(s+l') - \mathbf{X}_{v_{s+l'}})^{\mathrm{T}} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}},$$

$$L_5 = \sum_{s=1}^{n-l} \frac{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s+l) \sum_{l'=0}^{l-1} (e(\mathbf{J}(s+l')) - e(\mathbf{J}_{v_{s+l'}}))^{\mathrm{T}} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}},$$

$$L_6 = \sum_{s=1}^{n-l} \frac{\mathbf{p}_{ij}^{\mathrm{T}} (\mathbf{X}(s+l) - \mathbf{X}_{v_{s+l}}) \sum_{l'=0}^{l-1} (e(\mathbf{J}(s+l')) - e(\mathbf{J}_{v_{s+l'}}))^{\mathrm{T}} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}},$$

$$L_7 = \sum_{s=1}^{n-l} \frac{\mathbf{p}_{ij}^{\mathrm{T}} \sum_{l'=0}^{l-1} \boldsymbol{\epsilon}(s+l') (\mathbf{X}(s+l) - \mathbf{X}_{v_{s+l}})^{\mathrm{T}} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}},$$

$$L_8 = \sum_{s=1}^{n-l} \frac{\mathbf{p}_{ij}^{\mathrm{T}} \sum_{l'=0}^{l-1} \boldsymbol{\epsilon}(s+l') (e(\mathbf{J}(s+l)) - e(\mathbf{J}_{v_{s+l}}))^{\mathrm{T}} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}},$$

$$L_9 = \sum_{s=1}^{n-l} \frac{\mathbf{p}_{ij}^{\mathrm{T}} \sum_{l'=0}^{l-1} (\mathbf{X}(s+l') - \mathbf{X}_{v_{s+l'}}) (e(\mathbf{J}(s+l)) - e(\mathbf{J}_{v_{s+l}}))^{\mathrm{T}} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}}.$$

Consider L_1 first. Denoting $E_j(\cdot) = E(\cdot|\mathcal{F}_{-j})$, and $\mathrm{cov}_j(\cdot,\cdot)$ the corresponding covariance operator given $\mathcal{F}_{-j}$, we have

$$\begin{aligned} E_j(L_1) &= \sum_{s=1}^{n-l} O(1) \cdot \mathrm{cov}_j \left(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s+l), \mathbf{p}_{ij}^{\mathrm{T}} \sum_{l'=0}^{l-1} \boldsymbol{\epsilon}(s+l') \right) \\ &\leq \sum_{s=1}^{n-l} O(1) \cdot O(n^{-1}) (\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\epsilon,s+l}^{j} \mathbf{p}_{ij})^{1/2} O(l + l^2 n^{-1})^{1/2} = O(l^{1/2}), \end{aligned}$$

where we used Assumption (E3) on the weak correlation (at order n^{-1}) between functions of $\boldsymbol{\epsilon}(s_1)$ and $\boldsymbol{\epsilon}(s_2)$ when $s_1 \neq s_2$, and the boundedness of $\left\| \boldsymbol{\Sigma}_{\epsilon,s}^{j} \right\|$ in Assumption (E1). Also,

$$\begin{aligned} E_j(L_1^2) &= \sum_{s_1,s_2=1}^{n-l} O(1) \cdot E_j \left(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s_1+l) \mathbf{p}_{ij}^{\mathrm{T}} \sum_{l'=0}^{l-1} \boldsymbol{\epsilon}(s_1+l') \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s_2+l) \mathbf{p}_{ij}^{\mathrm{T}} \sum_{l'=0}^{l-1} \boldsymbol{\epsilon}(s_2+l') \right) \\ &= O(1) \cdot O(nl + n^2 \cdot n^{-1} \cdot l) = O(nl), \end{aligned}$$

so that

$$L_1 = O_P(n^{1/2}l^{1/2}).$$

To find the rate of L_2, define

$$\begin{aligned} A_d^{ij}(s) &= \sum_{m=1}^{p-b_s-1} \frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{D}_m^s\mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}})\mathbf{Z}_{d,s}^j(m)}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}}, \\ A_v^{ij}(s) &= \sum_{m=1}^{p-b_s-1} \frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{D}_m^s\boldsymbol{\Sigma}(t_s^{j_m}, t_s^{j_{m+1}})^{1/2}\mathbf{Z}_{v,s}^j(m)}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}}. \end{aligned} \tag{3.5}$$

Then we can decompose $L_2 = L_{d,d} + L_{d,v} + L_{v,d} + L_{v,v}$, where

$$L_{d,v} = \sum_{s=1}^{n-l} A_d^{ij}(s+l) \sum_{l'=0}^{l-1} A_v^{ij}(s+l'),\ L_{v,d} = \sum_{s=1}^{n-l} A_v^{ij}(s+l) \sum_{l'=0}^{l-1} A_d^{ij}(s+l'),$$

and the two other terms are defined similarly. Going through the proof of Lemma 2 in Lam and Feng (2018), we have

$$A_v^{ij}(s) = O_p(n^{-1/2}),\ \ A_d^{ij}(s) = \begin{cases} O_P(p_f^{1/2}n^{-1}), & \text{non-random drift;} \\ O_P(p_f^{1/2}p^{\alpha-1/2}n^{-1}), & \text{random drift.} \end{cases}$$

Hence for non-random drift, using Burkholder's inequality on all except $L_{d,d}$,

$$\begin{aligned} &L_{d,d} = O_P(p_f l n^{-1}),\ L_{d,v},\ L_{v,d} = O_P(p_f^{1/2} l n^{1/2} \cdot n^{-1/2} \cdot n^{-1}) = O_P(p_f^{1/2} l n^{-1}), \\ &L_{v,v} = O_P(l^{1/2} n^{-1/2}), \end{aligned}$$

implying that

$$L_2 = O_P(p_f l n^{-1} + l^{1/2} n^{-1/2}),$$

which in fact is true for the case of random drift as well.

Using Assumption (W1) to (W3), we have

$$\begin{aligned} L_3 = O_p(l n^{-1/2}), \quad L_5, L_8 = O_P(l n^{-1/4}), \\ L_6, L_9 = O_P(p_f^{1/2} \cdot l n^{-1/2} \cdot n^{-1/4}) = O_P(p_f^{1/2} l n^{-3/4}). \end{aligned}$$

Finally, using Assumption (E3) and the fact that $\{A_v^{ij}(s)\}$ is a sequence of indepen-

dent variables in the index s,

$$L_4,\ L_7 = O_P(n^{1/2} \cdot (p_f^{1/2} l n^{-1} + l^{1/2} n^{-1/2})) = O_P(p_f^{1/2} l n^{-1/2} + l^{1/2}).$$

Hence we have the result as stated, since $l^{1/2}n^{1/2}$ dominate all other term no matter $p_f = 1$ or $p_f = p \asymp n$. $\square$

3.7.1 Proof of Theorem 1

For NER-MSRVM in (3.3), $i = 1, \ldots, p$, $j = 1, \ldots, L$ with $\mathbf{P}_{-j} = (\mathbf{p}_{1j}, \ldots, \mathbf{p}_{pj})$, we can decompose

$$\begin{aligned}\mathbf{p}_{ij}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)^M \mathbf{p}_{ij} &= \mathbf{p}_{ij}^{\mathrm{T}} \sum_{m=1}^{M} a_m [\widetilde{\mathbf{Y}}, \widetilde{\mathbf{Y}}^{\mathrm{T}}]_j^{(K_m)} \mathbf{p}_{ij} + \mathbf{p}_{ij}^{\mathrm{T}} \zeta \left([\widetilde{\mathbf{Y}}, \widetilde{\mathbf{Y}}^{\mathrm{T}}]_j^{(K_1)} - [\widetilde{\mathbf{Y}}, \widetilde{\mathbf{Y}}^{\mathrm{T}}]_j^{(K_M)} \right) \mathbf{p}_{ij} \\ &= I_1 + 2I_2 + I_3,\end{aligned}$$

where $\widetilde{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)^M$ is the MSRVM in (3.4) constructed using jumps-removed data, and

$$\begin{aligned}I_1 &= \mathbf{p}_{ij}^{\mathrm{T}} \sum_{m=1}^{M} a_m [\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(K_m)} \mathbf{p}_{ij} + \mathbf{p}_{ij}^{T} \zeta \left([\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(K_1)} - [\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(K_M)} \right) \mathbf{p}_{ij} \\ &= \sum_{m=1}^{M} a_m \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(K_m)} \mathbf{p}_{ij} + \zeta \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(K_1)} \mathbf{p}_{ij} - \zeta \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(K_M)} \mathbf{p}_{ij}, \\ I_2 &= \sum_{m=1}^{M} a_m \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(K_m)} \mathbf{p}_{ij} + \zeta \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(K_1)} \mathbf{p}_{ij} - \zeta \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(K_M)} \mathbf{p}_{ij}, \\ I_3 &= \sum_{m=1}^{M} a_m \mathbf{p}_{ij}^{\mathrm{T}} [\mathbf{E}, \mathbf{E}^{\mathrm{T}}]_j^{(K_m)} \mathbf{p}_{ij} + \zeta \mathbf{p}_{ij}^{\mathrm{T}} [\mathbf{E}, \mathbf{E}^{\mathrm{T}}]_j^{(K_1)} \mathbf{p}_{ij} - \zeta \mathbf{p}_{ij}^{\mathrm{T}} [\mathbf{E}, \mathbf{E}^{\mathrm{T}}]_j^{(K_M)} \mathbf{p}_{ij}.\end{aligned}$$

Going through the proof of Lemma 3 in Lam and Feng (2018), we have

$$\max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \frac{\mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(K_m)} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} \right| = O_P(K_m^{-1/2} + n^{-1/4}). \tag{3.6}$$

By definition, $K_m = N + m$, with $N \asymp n^{2/3}$ and $M \asymp n^{1/2}$. Hence it is easy to see that all K_m has the same order at $n^{2/3}$ for $m = 1, \ldots, M$. Also, since $\sum_{m=1}^{M} a_m = 1$

and $\zeta \asymp n^{-1/6}$, it is easy to show that

$$\begin{aligned}
\left| \sum_{m=1}^{M} a_m \frac{\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(K_m)} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} - 1 \right| &= O_P(K_m^{1/2} n^{-1/2} + p_f^{1/2} n^{-1/2}) \\
&= O_P(n^{-1/6} + p_f^{1/2} n^{-1/2}), \\
\zeta \cdot \left(\left| \frac{\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(K_1)} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} \right| + \left| \frac{\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(K_M)} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} \right| \right) &= O_p(n^{-1/6} \cdot (K_m^{1/2} n^{-1/2} + p_f^{1/2} n^{-1/2} + 1)) \\
&= O_P(n^{-1/6}).
\end{aligned}$$

where both results above used Lemma 3. Since the above bounds are independent of the indices i and j, we have established that

$$\max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \frac{I_1}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} - 1 \right| = O_P(n^{-1/6} + p_f^{1/2} n^{-1/2}). \tag{3.7}$$

Similarly, to find the rate of I_2, using (3.6),

$$\begin{aligned}
\sum_{m=1}^{M} a_m \frac{\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(K_m)} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} &= O_P(K_m^{-1/2} + n^{-1/4}) = O_P(n^{-1/4}), \\
\zeta \left(\frac{\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(K_1)} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} - \frac{\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(K_M)} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} \right) &= O_P(n^{-1/6} \cdot (K_m^{-1/2} + n^{-1/4})) = O_P(n^{-5/12}).
\end{aligned}$$

Since the above two results are free of all indices i and j,

$$\max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \frac{I_2}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} \right| = O_P(n^{-1/4}). \tag{3.8}$$

To find the rate for I_3, consider the decomposition

$$\frac{\mathbf{p}_{ij}^{\mathrm{T}}[\mathbf{E},\mathbf{E}^{\mathrm{T}}]_j^{(K_m)}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = \sum_{\ell=1}^{3} I_{3,\ell} + 2\sum_{\ell=1}^{3} I_{3,\ell}, \text{ where}$$

$$I_{3,1}(m) = \frac{1}{K_m}\sum_{s,s-K_m\in S^j(K_m)} \frac{(\mathbf{p}_{ij}^{\mathrm{T}}(\boldsymbol{\epsilon}(s)-\boldsymbol{\epsilon}(s-K_m)))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

$$I_{3,2} = \frac{1}{K_m}\sum_{s,s-K_m\in S^j(K_m)} \frac{(\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s)-\mathbf{X}_{v_s}+\mathbf{X}_{v_{s-K_m}}-\mathbf{X}(s-K_m)))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

$$I_{3,3} = \frac{1}{K_m}\sum_{s,s-K_m\in S^j(K_m)} \frac{(\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}(s))-e(\mathbf{J}_{v_s})-e(\mathbf{J}(s-K_m))+e(\mathbf{J}_{v_{s-K_m}})))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

$$I_{3,4} = \frac{1}{K_m}\sum_{s,s-K_m\in S^j(K_m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\boldsymbol{\epsilon}(s)-\boldsymbol{\epsilon}(s-K_m))(\mathbf{X}(s)-\mathbf{X}_{v_s}+\mathbf{X}_{v_{s-K_m}}-\mathbf{X}(s-K_m))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

$$I_{3,5} = \frac{1}{K_m}\cdot \sum_{s,s-K_m\in S^j(K_m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\boldsymbol{\epsilon}(s)-\boldsymbol{\epsilon}(s-K_m))(e(\mathbf{J}(s))-e(\mathbf{J}_{v_s})-e(\mathbf{J}(s-K_m))+e(\mathbf{J}_{v_{s-K_m}}))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

$$I_{3,6} = \frac{1}{K_m}\sum_{s,s-K_m\in S^j(K_m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s)-\mathbf{X}_{v_s}+\mathbf{X}_{v_{s-K_m}}-\mathbf{X}(s-K_m))}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} \cdot (e(\mathbf{J}(s))-e(\mathbf{J}_{v_s})-e(\mathbf{J}(s-K_m))+e(\mathbf{J}_{v_{s-K_m}}))^{\mathrm{T}}\mathbf{p}_{ij}.$$

We consider $I_{3,2}$ first, which by Lemma 4 has

$$I_{3,2} = O_P(nK_m^{-1}\cdot n^{-1}) = O_P(K_m^{-1}) = O_P(n^{-2/3}).$$

Using Assumption (W1) to (W3) and the rate $n^{-1/4}$ of jumps removal in Fan and Wang (2007), we have

$$I_{3,3} = O_P(n^{-1/2}),\ I_{3,5} = O_P(n^{-1/4}),$$

$$I_{3,6} = O_P(n^{-1/2}n^{-1/4}) = O_P(n^{-3/4}).$$

Using Assumption (E3), we have

$$I_{3,4} = O_P(K_m^{-2}\cdot n\cdot n^{-1}) = O_P(n^{-2/3}).$$

Instead of considering the rate of $I_{3,1}(m)$, we consdier

$$\begin{aligned}
&\mathbf{p}_{ij}^T\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}(\sum_{m=1}^{M} a_m I_{3,1}(K_m)+\zeta(I_{3,1}(K_1)-I_{3,1}(K_M)))\\
=&\sum_{m=1}^{M} a_m\frac{1}{K_m}\sum_{s,s-K_m\in S^j(K_m)}(\mathbf{p}_{ij}^T(\boldsymbol{\epsilon}(s)-\boldsymbol{\epsilon}(s-K_m)))^2\\
&+\zeta(\frac{1}{K_1}\sum_{s,s-K_1\in S^j(K_1)}(\mathbf{p}_{ij}^T(\boldsymbol{\epsilon}(s)-\boldsymbol{\epsilon}(s-K_1)))^2 \qquad (3.9)\\
&-\frac{1}{K_M}\sum_{s,s-K_M\in S^j(K_M)}(\mathbf{p}_{ij}^T(\boldsymbol{\epsilon}(s)-\boldsymbol{\epsilon}(s-K_M)))^2)\\
=&J_1-2J_2+J_3, \qquad (3.10)
\end{aligned}$$

where

$$\begin{aligned}
J_1=&\sum_{m=1}^{M} a_m\frac{1}{K_m}\sum_{s,s-K_m\in S^j(K_m)}(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s))^2\\
&+\zeta\big(\frac{1}{K_1}\sum_{s,s-K_1\in S^j(K_1)}(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s))^2-\frac{1}{K_M}\sum_{s,s-K_M\in S^j(K_M)}(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s))^2\big),\\
J_2=&\sum_{m=1}^{M} a_m\frac{1}{K_m}\sum_{s,s-K_m\in S^j(K_m)}(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s)\boldsymbol{\epsilon}(s-K_m)^{\mathrm{T}}\mathbf{p}_{ij})\\
&+\zeta\big(\frac{1}{K_1}\sum_{s,s-K_1\in S^j(K_1)}(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s)\boldsymbol{\epsilon}(s-K_1)^{\mathrm{T}}\mathbf{p}_{ij})\\
&-\frac{1}{K_M}\sum_{s,s-K_M\in S^j(K_M)}(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s)\boldsymbol{\epsilon}(s-K_M)^{\mathrm{T}}\mathbf{p}_{ij})\big),\\
J_3=&\sum_{m=1}^{M} a_m\frac{1}{K_m}\sum_{s,s-K_m\in S^j(K_m)}(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s-K_m))^2\\
&+\zeta\big(\frac{1}{K_1}\sum_{s,s-K_1\in S^j(K_1)}(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s-K_1))^2-\frac{1}{K_M}\sum_{s,s-K_M\in S^j(K_M)}(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s-K_M))^2\big).
\end{aligned}$$

Writing $g_{m,s}^{ij}=\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s)\boldsymbol{\epsilon}(s-K_m)^{\mathrm{T}}\mathbf{p}_{ij}$, by Lemma 2.7 of Bai and Silverstein (1998),

$$E\Big\{\Big(\frac{1}{K_m}\sum_{s,s-K_m\in S^j(m)} g_{m,s}^{ij}\Big)^2\Big|\mathcal{F}_{ij}\Big\}=O(K_m^{-2}n\cdot 1+n^{-1}\cdot K_m^{-2}n^2\cdot 1)=O(K_m^{-2}n),\ \text{hence}$$

$$\frac{1}{K_m}\sum_{s,s-K_m\in S^j(m)} g_{m,s}^{ij}=O_P(K_m^{-1}n^{1/2})=O_P(n^{-1/6}),$$

which implies that

$$J_2 = O_P(n^{-1/6} + n^{-1/3}) = O_P(n^{-1/6}).$$

We can further decompose $J_1 = J_{11} + J_{12} + J_{13}$, where

$$\begin{aligned}
J_{11} &= \sum_{m=1}^{M} a_m \frac{1}{K_m} \sum_{s, s-K_m \in S^j(K_m)} \big((\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s))^2 - \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\epsilon,s}^j \mathbf{p}_{ij}\big), \\
J_{12} &= \zeta\Bigg(\frac{1}{K_1} \sum_{s, s-K_1 \in S^j(K_1)} \big((\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s))^2 - \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\epsilon,s}^j \mathbf{p}_{ij}\big) \\
&\quad - \frac{1}{K_M} \sum_{s, s-K_M \in S^j(K_M)} \big((\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s))^2 - \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\epsilon,s}^j \mathbf{p}_{ij}\big) \Bigg), \\
J_{13} &= \sum_{m=1}^{M} a_m \frac{1}{K_m} \sum_{s, s-K_m \in S^j(K_m)} \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon},s}^j \mathbf{p}_{ij} \\
&\quad + \zeta\big(\frac{1}{K_1} \sum_{s, s-K_1 \in S^j(K_1)} \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon},s}^j \mathbf{p}_{ij} - \frac{1}{K_M} \sum_{s, s-K_M \in S^j(K_M)} \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon},s}^j \mathbf{p}_{ij} \big).
\end{aligned}$$

Now define $g_s^{ij} = \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s) - \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\epsilon,s}^j \mathbf{p}_{ij}$. Using Lemma 2.7 of Bai and Silverstein (1998) under Assumption (E1) to (E3), we have

$$\begin{aligned}
&E_j\big(\big(\frac{1}{K_m} \sum_{s, s-K_m \in S^j(K_m)} g_s^{ij}\big)^2 | \{\boldsymbol{\Sigma}_{\epsilon,u}, u \in [0,1]\}\big) \\
&= K_m^{-2} \sum_{s, s-K_m \in S^j(K)} E_j\big((g_s^{ij})^2 | \{\boldsymbol{\Sigma}_{\epsilon,u}, u \in [0,1]\}\big) \\
&\quad + K_m^{-2} \sum_{s_1 \neq s_2} E_j\big(g_{s_1}^{ij} g_{s_2}^{ij} | \{\boldsymbol{\Sigma}_{\epsilon,u}, u \in [0,1]\}\big) \\
&= O(K_m^{-2} n \cdot 1 + K_m^{-2} n^2 \cdot n^{-1} \cdot 1) = O_P(n^{-1/3}).
\end{aligned}$$

The above implies that

$$J_{11} = O_P(n^{-1/6}) = J_{12}.$$

As for the rate of J_{13}, we can do a further decomposition $J_{13} = J_{13}^{(1)} + J_{13}^{(2)} + J_{13}^{(3)}$,

where $c_m = \frac{n-K_m+1}{K_m}$ and

$$
\begin{aligned}
J_{13}^{(1)} =& \sum_{m=1}^{M} a_m \big(\frac{1}{K_m} \sum_{s, s-K_m \in S^j(K_m)} \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon},s}^{j} \mathbf{p}_{ij} - c_m \mathbf{p}_{ij}^{\mathrm{T}} E(\boldsymbol{\Sigma}_{\epsilon,s}^{j}) \mathbf{p}_{ij} \big), \\
J_{13}^{(2)} =& \zeta \Bigg(\big(\frac{1}{K_1} \sum_{s, s-K_1 \in S^j(K_1)} \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon},s}^{j} \mathbf{p}_{ij} - c_1 \mathbf{p}_{ij}^{\mathrm{T}} E(\boldsymbol{\Sigma}_{\epsilon,s}^{j}) \mathbf{p}_{ij} \big) \\
& - \big(\frac{1}{K_M} \sum_{s, s-K_M \in S^j(K_M)} \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon},s}^{j} \mathbf{p}_{ij} - c_M \mathbf{p}_{ij}^{\mathrm{T}} E(\boldsymbol{\Sigma}_{\epsilon,s}^{j}) \mathbf{p}_{ij} \big) \Bigg), \\
J_{13}^{(3)} =& \Bigg(\sum_{m=1}^{M} a_m c_m + \zeta \big(c_1 - c_M \big) \Bigg) \mathbf{p}_{ij}^{\mathrm{T}} E(\boldsymbol{\Sigma}_{\epsilon,s}^{j}) \mathbf{p}_{ij}.
\end{aligned}
$$

Parallel to the argument used for J_{11}, under Assumption (E3), we have

$$J_{13}^{(1)} = O_P(K_m^{-2} n + K_m^{-2} n^2 \cdot n^{-1})^{1/2} = O_P(n^{-1/6}).$$

Similarly,

$$J_{13}^{(2)} = O_P(n^{-1/6} \cdot n^{-1/6}) = O_P(n^{-1/3}).$$

By the same technique used in the proof of Theorem 1 in Tao et al. (2013), and the definition of a_m, ζ in 3.3 and c_m shown above, we have $J_{13}^{(3)} = 0$. Hence

$$J_{13} = O_P(n^{-1/6}).$$

This implies that

$$J_1 = O_P(n^{-1/6}) = J_3.$$

Combining these rates, we have

$$\sum_{m=1}^{M} a_m I_{3,1}(K_m) + \zeta(I_{3,1}(K_1) - I_{3,1}(K_M)) = O_P(n^{-1/6}).$$

Hence this, together with the rates for $I_{3,2}$ to $I_{3,6}$, allow us to conclude that

$$\max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \frac{I_3}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} \right| = O_P(n^{-1/6}). \tag{3.11}$$

Combining (3.7), (3.8) and (3.11), we thus have

$$\max_{\substack{i=1,\ldots,p\\ j=1,\ldots,L}} \left| \frac{\mathbf{p}_{ij}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)^M \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} - 1 \right| = O_P(n^{-1/6} + p_f^{1/2} n^{-1/2}).$$

For $\widehat{\mathbf{\Sigma}}(0,1)^M$, we have

$$\begin{aligned}
&\left\|\widehat{\mathbf{\Sigma}}(0,1)^M \mathbf{\Sigma}_{\mathrm{Ideal}}(0,1)^{-1} - \mathbf{I}_p\right\| \\
&= \left\| \sum_{j=1}^{L} (\widehat{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)^M \mathbf{\Sigma}_{\mathrm{Ideal}}(\tau_{j-1},\tau_j)^{-1} - \mathbf{I}_p) \mathbf{\Sigma}_{\mathrm{Ideal}}(\tau_{j-1},\tau_j) \mathbf{\Sigma}_{\mathrm{Ideal}}(0,1)^{-1} \right\| \\
&\le \sum_{j=1}^{L} \left\|\widehat{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)^M \mathbf{\Sigma}_{\mathrm{Ideal}}(\tau_{j-1},\tau_j)^{-1} - \mathbf{I}_p\right\| \\
&\quad \cdot \left\| \mathrm{diag}(\mathbf{P}_{-j}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{P}_{-j}) \cdot \left(\mathrm{diag}(\mathbf{P}_{-j}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{P}_{-j}) + \sum_{i\neq j} \mathbf{P}_{-j}^{\mathrm{T}} \mathbf{\Sigma}_{\mathrm{Ideal}}(\tau_{i-1},\tau_i)\mathbf{P}_{-j} \right)^{-1} \right\| \\
&= O_P\Bigg((n^{-1/6} + p_f^{1/2} n^{-1/2}) \\
&\qquad \cdot \max_{j=1,\ldots,L} \left\| \left(\mathbf{I}_p + \sum_{i\neq j} \mathbf{P}_{-j}^{\mathrm{T}} \mathbf{\Sigma}_{\mathrm{Ideal}}(\tau_{i-1},\tau_i)\mathbf{P}_{-j} \mathrm{diag}^{-1}(\mathbf{P}_{-j}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{P}_{-j}) \right)^{-1} \right\| \Bigg) \\
&= O_P(n^{-1/6} + p_f^{1/2} n^{-1/2}). \ \square
\end{aligned}$$

3.7.2 Proof of the Theorem 2

In this proof we drop the subscripts j in $\boldsymbol{\gamma}_j^{(h)}(\widetilde{\mathbf{Y}}_j^{(J)})$ since we are working within partition j already. Define, for a time series of vectors $\mathbf{Y}(\cdot)$,

$$\begin{aligned}
&\mathbf{Y}^{(J)}(1) = J^{-1} \sum_{l=0}^{J-1} \mathbf{Y}(l), \quad \mathbf{Y}^{(J)}(n-2J+1) = J^{-1} \sum_{l=n-J+1}^{n} \mathbf{Y}(l), \\
&\mathbf{Y}^{(J)}(s) = \widetilde{\mathbf{Y}}(s+J-1),\ s = 2,\ldots,n-2J.
\end{aligned}$$

To extract the jittering effects for analysis, define

$$\begin{aligned}
&\widetilde{\mathbf{Y}}^{*(J)}(1) = \widetilde{\mathbf{X}}(J), \quad \widetilde{\mathbf{Y}}^{*(J)}(n-2J+1) = \widetilde{\mathbf{X}}(n-J), \\
&\widetilde{\mathbf{Y}}^{*(J)}(s) = \widetilde{\mathbf{Y}}^{(J)}(s) = \widetilde{\mathbf{Y}}(s+J-1),\ s = 2,\ldots,n-2J.
\end{aligned}$$

Then we have

$$\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{K}(\widetilde{\mathbf{Y}}^{(J)})\mathbf{p}_{ij} - \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{K}(\widetilde{\mathbf{Y}}^{*(J)})\mathbf{p}_{ij} = \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}_J^{(0)}\mathbf{p}_{ij} + 2\sum_{h=1}^{n-2J} k_H(h)\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}_J^{(h)}\mathbf{p}_{ij},$$

where $k_H(h) = k((h-1)/H)$ or $k(h/H)$, and $\mathbf{E}_J^{(h)} = \mathbf{E}_{J,1}^{(h)} + \mathbf{E}_{J,2}^{(h)}$, where for $h > 0$,

$$\begin{aligned}
\mathbf{E}_{J,1}^{(h)} &= (\widetilde{\mathbf{Y}}^{(J)}(h+2) - \widetilde{\mathbf{Y}}^{(J)}(h+1))(\widetilde{\mathbf{X}}(J) - \widetilde{\mathbf{Y}}^{(J)}(1))^{\mathrm{T}},\\
\mathbf{E}_{J,2}^{(h)} &= (\widetilde{\mathbf{Y}}^{(J)}(n-2J+1) - \widetilde{\mathbf{X}}(n-J))(\widetilde{\mathbf{Y}}^{(J)}(n-2J+1-h) - \widetilde{\mathbf{Y}}^{(J)}(n-2J-h))^{\mathrm{T}},\\
\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}_{J,1}^{(0)}\mathbf{p}_{ij} &= (\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{Y}}^{(J)}(2) - \widetilde{\mathbf{Y}}^{(J)}(1)))^2 - (\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{Y}}^{(J)}(2) - \widetilde{\mathbf{X}}(J)))^2,\\
\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}_{J,2}^{(0)}\mathbf{p}_{ij} &= (\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{Y}}^{(J)}(n-2J+1) - \widetilde{\mathbf{Y}}^{(J)}(n-2J)))^2 - (\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}(n-J) - \widetilde{\mathbf{Y}}^{(J)}(n-2J)))^2.
\end{aligned}$$

We now present a proposition on the rate of jittering effects from NER-KRVM or NER-KRPVM.

Proposition 1. *Let all the assumptions in Theorem 4 hold. Then with $p/n \to c \geq 0$,*

$$\begin{aligned}
&\frac{\mathbf{p}_{ij}^{T}\mathbf{E}_J^{(0)}\mathbf{p}_{ij} + 2\sum_{h=1}^{n-2J} k_H(h)\mathbf{p}_{ij}^{T}\mathbf{E}_J^{(h)}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{T}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\\
&= O_P(J^{-1} + Jn^{-1} + (J^{-1/2} + J^{1/2}n^{-1/2})(H^{1/2}n^{-1/2} + p_f^{1/2}Hn^{-1} + n^{-1/4})),
\end{aligned}$$

where $p_f = 1$ when there are no factors, or $p_f = p$ when there are pervasive factors in the log-price processes.

Proof of Proposition 1. We shall only consider $E_{J,1}^{(h)}$ for $h \geq 0$ in this proof since $E_{J,2}^{(h)}$ can be treated exactly the same with the same rate. Consider

$$\begin{aligned}
&\frac{\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{Y}}^{(J)}(1) - \widetilde{\mathbf{X}}(J))}{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}} = I_1 + I_2 + I_3, \text{ where}\\
&I_1 = J^{-1}\sum_{s=0}^{J-1}\frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s) - \mathbf{X}(J))}{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}}, \quad I_2 = J^{-1}\sum_{s=0}^{J-1}\frac{\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}(s)) - e(\mathbf{J}(J)))}{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}},\\
&I_3 = J^{-1}\sum_{s=0}^{J-1}\frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s)}{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}}.
\end{aligned}$$

Under Assumption (W1) to (W3) and the rate of jumps removal at $n^{-1/4}$ from Fan and Wang (2007),

$$I_2 = O_P(n^{-1/4}).$$

Also, with Assumption (E3),

$$E_j(I_3)^2 = O(J^{-2} \cdot J + J^{-2} \cdot J^2 \cdot n^{-1}) = O(J^{-1}),$$

so that

$$I_3 = O_P(J^{-1/2}).$$

For I_1, using the result of Lemma 4 and Assumptions (D1) and (V1),

$$\begin{aligned} I_1 &= J^{-1} \sum_{s=0}^{J-1} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s) - \mathbf{X}_{v_s})}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij})^{1/2}} + J^{-1} \sum_{s=0}^{J-1} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}_{v_s} - \mathbf{X}_{v_J})}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij})^{1/2}} \\ &\quad + J^{-1} \sum_{s=0}^{J-1} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}_{v_J} - \mathbf{X}(J))}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij})^{1/2}} \\ &= O_P(n^{-1/2}) + O_P(J^{-1} \sum_{s=0}^{J-1} (p_f^{1/2}(J-s)^{1/2} n^{-1} + (J-s)^{1/2} n^{-1/2})) + O_P(n^{-1/2}) \\ &= O_P(p_f^{1/2} J^{1/2} n^{-1} + J^{1/2} n^{-1/2}) = O_P(J^{1/2} n^{-1/2}). \end{aligned}$$

Hence we have

$$\frac{\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{Y}}^{(J)}(1) - \widetilde{\mathbf{X}}(J))}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij})^{1/2}} = O_P(J^{-1/2} + J^{1/2} n^{-1/2} + n^{-1/4}). \tag{3.12}$$

Now consider

$$\begin{aligned} JE &= \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{E}_{J,1}^{(h)} \mathbf{p}_{ij} + 2 \sum_{h=1}^{n-2J} k_H(h) \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{E}_{J,1}^{(h)} \mathbf{p}_{ij} \\ &= [\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{Y}}^{(J)}(1) - \widetilde{\mathbf{X}}(J))]^2 + 2\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{Y}}^{(J)}(1) - \widetilde{\mathbf{X}}(J)) \cdot R(J), \quad \text{where} \\ R(J) &= \mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}(J) - \widetilde{\mathbf{Y}}(J+1)) + \sum_{h=1}^{n-2J} k_H(h) \mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{Y}}(h+J) - \widetilde{\mathbf{Y}}(h+J+1)). \end{aligned}$$

If $k_H(h) = k((h-1)/H)$, then

$$
\begin{aligned}
&R(J) \\
&= \mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{X}}(J) + \sum_{h=2}^{n-2J}(k_H(h) - k_H(h-1))\mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{Y}}(h+J) + k_H(n-2J)\mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{Y}}(n-J+1) \\
&= \sum_{h=2}^{n-2J}(k_H(h) - k_H(h-1))\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{Y}}(h+J) - \widetilde{\mathbf{X}}(J)) + k_H(n-2J)\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}(J) + \widetilde{\mathbf{Y}}(n-J+1)) \\
&= \sum_{h=2}^{n-2J}k'(\xi_{h-1})H^{-1}(\widetilde{\mathbf{Y}}(h+J) - \widetilde{\mathbf{X}}(J)) + O_P(n^{-1}),
\end{aligned}
$$

where $(h-2)/H \leq \xi_{h-1} \leq (h-1)/H$. If $k_H(h) = k(h/H)$, following the above steps, we can show

$$
\begin{aligned}
R(J) &= \sum_{h=1}^{n-2J}k'(\xi_h)H^{-1}(\widetilde{\mathbf{Y}}(h+J) - \widetilde{\mathbf{X}}(J)) + O_P(n^{-1}) \\
&= D_1 + D_2 + D_3 + O_P(1), \quad \text{where} \\
D_1 &= \sum_{h=1}^{n-2J}k'(\xi_h)H^{-1}\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(h+J) - \mathbf{X}(J)), \\
D_2 &= \sum_{h=1}^{n-2J}k'(\xi_h)H^{-1}\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}(h+J)) - e(\mathbf{J}(J))), \\
D_3 &= \sum_{h=1}^{n-2J}k'(\xi_h)H^{-1}\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(h+J).
\end{aligned}
$$

Using Assumption (W1) to (W3) and the rate of jump removal in Fan and Wang (2007),

$$D_2/(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2} = O_P(n^{-1/4}).$$

With Assumption (E3),

$$D_3/(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2} = O_P(H^{-1/2} + n^{-1/2}) = O_P(n^{-1/2}).$$

We further decompose $D_1 = D_{11} + D_{12} + D_{13}$, where

$$
\begin{aligned}
D_{11} &= \sum_{h=1}^{n-2J} k'(\xi_h) H^{-1} \mathbf{p}_{ij}^{\mathrm{T}} (\mathbf{X}(h+J) - \mathbf{X}_{v_{h+J}}), \\
D_{12} &= \sum_{h=1}^{n-2J} k'(\xi_h) H^{-1} \mathbf{p}_{ij}^{\mathrm{T}} (\mathbf{X}_{v_{h+J}} - \mathbf{X}_{v_J}), \\
D_{13} &= \sum_{h=1}^{n-2J} k'(\xi_h) H^{-1} \mathbf{p}_{ij}^{\mathrm{T}} (\mathbf{X}_{v_J} - \mathbf{X}(J)).
\end{aligned}
$$

Using (3.5) and Assumption (A4), we have

$$
\begin{aligned}
\frac{D_{11}}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}} &= \sum_{h=1}^{n-2J} k'(\xi_h) H^{-1} (A_d^{ij}(h+J) + A_v^{ij}(h+J)) \\
&= O_P(H^{-1/2} n^{-1/2} + p_f^{1/2} n^{-1}).
\end{aligned}
$$

Using Lemma 4, we also have

$$
\frac{D_{13}}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}} = O_P(n^{-1/2}).
$$

To find the rate of D_{12}, We decompose $D_{12} = D_{12,d} + D_{12,v}$, where

$$
\begin{aligned}
\frac{D_{12,d}}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}} &= \sum_{h=1}^{n-2J} k'(e_h) H^{-1} \sum_{s=1}^{h} \frac{\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s+j-1}, v_{s+J}) \mathbf{Z}_{d,s}^j}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}} = O_P(p_f^{1/2} H n^{-1}), \\
\frac{D_{12,v}}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}} &= \sum_{h=1}^{n-2J} k'(e_h) H^{-1} \sum_{s=1}^{h} \frac{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(v_{s+j-1}, v_{s+J})^{1/2} \mathbf{Z}_{v,s}^j}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}} \\
&= \sum_{h=1}^{n-2J} k'(e_h) H^{-1} O_P(h^{1/2} n^{-1/2}) = O_P(H^{1/2} n^{-1/2}).
\end{aligned}
$$

Hence combining the rates obtained,

$$
\frac{R(J)}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}} = O_P(H^{1/2} n^{-1/2} + p_f^{1/2} H n^{-1} + n^{-1/4}). \tag{3.13}
$$

With (3.12) and (3.13), and the fact that the terms involving $\mathbf{E}_{J,2}^{(h)}$ give exactly the

same rates as those involving $\mathbf{E}_{J,1}^{(h)}$, the jittering effect has

$$\frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}_J^{(h)}\mathbf{p}_{ij} + 2\sum_{h=1}^{n-2J} k_H(h)\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}_J^{(h)}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}$$
$$= O_P(J^{-1} + Jn^{-1} + n^{-1/4} + (J^{-1/2} + J^{1/2}n^{-1/2})(H^{1/2}n^{-1/2} + p_f^{1/2}Hn^{-1} + n^{-1/4})). \ \square$$

To finish the proof of Theorem 2, using the notation in the proof of Proposition 1, we decompose

$$\frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{K}(\widetilde{\mathbf{Y}}^{*(J)})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} - 1 = K_1 + K_{21} + K_{22} + K_3, \text{ with}$$
$$K_1 = \frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}^{(0)}(\widetilde{\mathbf{X}}_v^{*(J)})\mathbf{p}_{ij} + 2\sum_{h=1}^{n-2J} k_H(h)\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}^{(h)}(\widetilde{\mathbf{X}}_v^{*(J)})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} - 1,$$
$$K_{21} = \frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}^{(0)}(\widetilde{\mathbf{X}}_v^{*(J)},\mathbf{E}^{*(J)})\mathbf{p}_{ij} + 2\sum_{h=1}^{n-2J} k_H(h)\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}^{(h)}(\widetilde{\mathbf{X}}_v^{*(J)},\mathbf{E}^{*(J)})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$
$$K_{22} = \frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}^{(0)}(\mathbf{E}^{*(J)},\widetilde{\mathbf{X}}_v^{*(J)})\mathbf{p}_{ij} + 2\sum_{h=1}^{n-2J} k_H(h)\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}^{(h)}(\mathbf{E}^{*(J)},\widetilde{\mathbf{X}}_v^{*(J)})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$
$$K_3 = \frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}^{(0)}(\mathbf{E}^{*(J)})\mathbf{p}_{ij} + 2\sum_{h=1}^{n-2J} k_H(h)\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}^{(h)}(\mathbf{E}^{*(J)})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}},$$

where we define

$$\widetilde{\mathbf{X}}_{v_s}^{*(J)} = \mathbf{X}_{v_{s+J-1}} + e(\mathbf{J}(s+J-1)),$$
$$\mathbf{E}^{*(J)}(s) = \begin{cases} \mathbf{0}, & s = 1, n-2J+1; \\ (\mathbf{X}(s+J-1) - \mathbf{X}_{v_{s+J-1}}) + \boldsymbol{\epsilon}(s+J-1), & \text{otherwise.} \end{cases},$$

and

$$\boldsymbol{\gamma}^{(h)}(\widetilde{\mathbf{X}}_v^{*(J)},\mathbf{E}^{*(J)}) = \sum_{s=h+2}^{n-2J+1}(\widetilde{\mathbf{X}}_{v_s}^{*(J)} - \widetilde{\mathbf{X}}_{v_{s-1}}^{*(J)})(\mathbf{E}^{*(J)}(s-h) - \mathbf{E}^{*(J)}(s-h-1))^{\mathrm{T}},$$

with similar definition for $\boldsymbol{\gamma}^{(h)}(\mathbf{E}^{*(J)},\widetilde{\mathbf{X}}_v^{*(J)})$.

Proposition 2. *Let all the assumptions in Theorem 4 hold. Then with $p/n \to c \geq 0$,*

$$K_1 = O_P(p_f^{1/2}H^{1/2}n^{-1/2} + p_f^{1/2}n^{-1/4}),$$

where $p_f = 1$ when there are no factors, or $p_f = p$ when there are pervasive factors in the log-price processes.

Proof of Proposition 2. Firstly, by Lemma 3, since $J = o(n)$, we immediately have

$$\left| \frac{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\gamma}^{(0)} (\widetilde{\mathbf{X}}_v^{*(J)}) \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} - 1 \right| = O_P(1 \cdot n^{-1/2} + p_f^{1/2} n^{-1/2}) = O_P(p_f^{1/2} n^{-1/2}).$$

Also, defining

$$z_{d,s}^{ij} = \frac{\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s+J-2}, v_{s+J-1}) \mathbf{Z}_{d,s+J-1}^{j}}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}}, \quad z_{v,s}^{ij} = \frac{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(v_{s+J-2}, v_{s+J-1})^{1/2} \mathbf{Z}_{v,s+J-1}^{j}}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}},$$
$$e_s^{ij} = \frac{\mathbf{p}_{ij}^{\mathrm{T}} (e(\mathbf{J}(s+J-1)) - e(\mathbf{J}(s+J-2)))}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}},$$

we can decompose (we omit any superscripts ij if no ambiguity arises)

$$\frac{\sum_{h=1}^{n-2J} k_H(h) \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\gamma}^{(h)} (\widetilde{\mathbf{X}}_v^{*(J)}) \mathbf{p}_{ij}}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}} = \sum_{r=1}^{7} M_r, \text{ where}$$
$$M_1 = \sum_{s=3}^{n-2J+1} z_{d,s} \sum_{h=1}^{s-2} k_H(h) z_{d,s-h}, \; M_2 = \sum_{s=3}^{n-2J+1} z_{d,s} \sum_{h=1}^{s-2} k_H(h) z_{v,s-h},$$
$$M_3 = \sum_{s=3}^{n-2J+1} z_{v,s} \sum_{h=1}^{s-2} k_H(h) z_{d,s-h}, \; M_4 = \sum_{s=3}^{n-2J+1} z_{v,s} \sum_{h=1}^{s-2} k_H(h) z_{v,s-h},$$
$$M_5 = \sum_{s=3}^{n-2J+1} e_s \sum_{h=1}^{s-2} k_H(h) (z_{d,s-h} + z_{v,s-h}),$$
$$M_6 = \sum_{s=3}^{n-2J+1} (z_{d,s} + z_{v,s}) \sum_{h=1}^{s-2} k_H(h) e_{s-h}, \; M_7 = \sum_{s=3}^{n-2J+1} e_s \sum_{h=1}^{s-2} k_H(h) e_{s-h}.$$

Using Assumption (D1), we have

$$M_1 = O_P(p_f H n^{-1}).$$

Using Assumption (D1) and (V1),

$$M_2 = O_P(p_f^{1/2} H^{1/2} n^{-1/2}).$$

Using the Burkholder's inequality,

$$M_3 = O_P(n \cdot n^{-1} \cdot p_f H^2 n^{-2})^{1/2} = O_P(p_f^{1/2} H n^{-1}).$$

Using the Burkholder's inequality again,

$$M_4 = O_P(n \cdot n^{-1} \cdot H n^{-1})^{1/2} = O_P(H^{1/2} n^{-1/2}).$$

Finally, with Assumption (W1) to (W3), using the jump removal rate $n^{-1/4}$ from Fan and Wang (2007),

$$M_5 = O_P(n^{-1/4} \cdot (p_f^{1/2} H n^{-1} + H^{1/2} n^{-1/2})),\ M_6 = O_P(p_f^{1/2} n^{-1/4} + n^{-1/4}),\ M_7 = O_P(n^{-1/2}).$$

Combining all the rates calculated, we can conclude that

$$K_1 = O_P(p_f^{1/2} H^{1/2} n^{-1/2} + p_f^{1/2} n^{-1/4}).\ \square$$

Proposition 3. *Let all the assumptions in Theorem 4 hold. Then with $p/n \to c \geq 0$,*

$$K_3 = \begin{cases} O_P(H^{-3/2} n^{1/2}), & k_H(h) = k((h-1)/H); \\ O_P(H^{-2} n), & k_H(h) = k(h/H). \end{cases}$$

Proof of Proposition 3. Firstly, define

$$r_s^{ij} = \frac{\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{E}^{*(J)}(s)}{(\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}},$$

so that by the definition of $\mathbf{E}^{*(J)}(s)$, $r_1^{ij} = r_{n-2J+1}^{ij} = 0$.

If no ambiguity arises, we drop the superscript in r_s^{ij} for easier presentation. Consider

the decomposition

$$
\begin{aligned}
&\frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}^{(0)}(\mathbf{E}^{*(J)})\mathbf{p}_{ij} + 2\sum_{h=1}^{n-2J+1} k_H(h)\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}^{(h)}(\mathbf{E}^{*(J)})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} \\
&= \sum_{s=2}^{n-2J+1}(r_s - r_{s-1})^2 + 2\sum_{h=1}^{n-2J} k_H(h) \sum_{s=h+2}^{n-2J+1}(r_s - r_{s-1})(r_{s-h} - r_{s-h-1}) \\
&= \sum_{s=2}^{n-2J+1}(r_s - r_{s-1})^2 + 2\sum_{s=2}^{n-2J+1}(r_s - r_{s-1})\sum_{h=1}^{s-2} k_H(h)(r_{s-h} - r_{s-h-1}) \\
&= \sum_{s=2}^{n-2J+1}(r_s - r_{s-1})\Big(r_s - r_{s-1} + 2\sum_{h=1}^{s-2} k_H(h)(r_{s-h} - r_{s-h-1})\Big) \\
&= r_2^2 + \sum_{s=3}^{n-2J+1}(r_s - r_{s-1})\Big((r_s + r_{s-1}) + (2k_H(1)r_{s-1} - 2r_{s-1}) \\
&\qquad + 2\sum_{h=2}^{s-2}\big(k_H(h) - k_H(h-1)\big)r_{s-h}\Big) \\
&= IE_1 + IE_2 + IE_3, \text{ where} \\
IE_1 &= \sum_{s=2}^{n-2J+1}(r_s^2 - r_{s-1}^2), \quad IE_2 = 2(k_H(1) - 1)\sum_{s=3}^{n-2J+1}(r_s - r_{s-1})r_{s-1}, \\
IE_3 &= 2\sum_{s=4}^{n-2J+1}(r_s - r_{s-1})\sum_{h=2}^{s-2} k'(\xi_h)H^{-1}r_{s-h},
\end{aligned}
$$

where $(h-2)/H \le \xi_h \le (h-1)/H$ if $k_H(h) = k((h-1)/H)$, and $(h-1)/H \le \xi_h \le h/H$ if $k_H(h) = k(h/H)$.

Firstly, since $r_1 = r_{n-2J+1} = 0$, we have

$$IE_1 = 0.$$

Secondly, $IE_2 = 0$ if $k_H(h) = k((h-1)/H)$ since then $k_H(1) = k(0) = 1$.

For $k_H(h) = k(h/H)$, using the fact that $k'(0) = 0$,

$$
\begin{aligned}
IE_2 &= 2k'(\xi_1)H^{-1}\sum_{s=3}^{n-2J+1}(r_s - r_{s-1})r_{s-1} \\
&= 2k''(\xi_1')\xi_1 H^{-1}\sum_{s=3}^{n-2J+1}(r_s - r_{s-1})r_{s-1}, \ 0 \le \xi_1' \le \xi_1 \le H^{-1}.
\end{aligned}
$$

Decompose further $r_s = r_{x,s} + r_{\epsilon,s}$, where

$$r_{x,s} = \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s) - \mathbf{X}_{v_s})}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}}, \quad r_{\epsilon,s} = \frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s)}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}}.$$

Then by Lemma 4 and Assumption (E1), we have

$$IE_2 \le 8H^{-2}k''(\xi_1') \sum_{s=3}^{n-2J+1} (r_{x,s-1}^2 + r_{\epsilon,s-1}^2) = O_P(H^{-2}(1+n)) = O_P(H^{-2}n).$$

Hence we have

$$IE_2 = \begin{cases} 0, & k_H(h) = k((h-1)/H); \\ O_P(H^{-2}n), & k_H(h) = k(h/H). \end{cases}$$

We decompose

$$\begin{aligned} IE_3 &= 2\sum_{s=4}^{n-2J+1} r_s \sum_{h=2}^{s-2} k'(\xi_h)H^{-1}r_{s-h} - 2\sum_{s=3}^{n-2J} r_s \sum_{h=2}^{s-1} k'(\xi_h)H^{-1}r_{s-h+1} \\ &= 2\sum_{s=3}^{n-2J} r_s \sum_{h=2}^{s-1} k'(\xi_h)H^{-1}(r_{s-h} - r_{s-h+1}) \\ &= 2\sum_{s=3}^{n-2J} r_s \left(\sum_{h=2}^{s-1} k'(\xi_h)H^{-1}r_{s-h} - \sum_{h=1}^{s-2} k'(\xi_{h+1})H^{-1}r_{s-h} \right) \\ &= 2\sum_{s=3}^{n-2J} r_s \left(\sum_{h=1}^{s-2} k''(\xi_h')(\xi_h - \xi_{h+1})H^{-1}r_{s-h} - k'(\xi_1)H^{-1}r_{s-1} \right), \ \xi_h \le \xi_h' \le \xi_{h+1} \\ &= IE_{3,1} + IE_{3,2}, \text{ where} \end{aligned}$$

$$IE_{3,1} = -2\sum_{s=3}^{n-2J} k'(\xi_1)H^{-1}r_s r_{s-1}, \quad IE_{3,2} = 2\sum_{s=3}^{n-2J} r_s \sum_{h=1}^{s-2} k''(\xi_h')(\xi_h - \xi_{h+1})H^{-1}r_{s-h}.$$

For $k_H(h) = k((h-1)/H)$, $\xi_1 = 0$ with $k'(\xi_1) = k_H(1) - 1 = 0$, and hence $IE_{3,1} = 0$.
For $k_H(h) = k(h/H)$, using Lemma 4 and Assumption (E3),

$$\begin{aligned} IE_{3,1} &= -2\sum_{s=3}^{n-2J} k''(\xi_1')\xi_1 H^{-1}r_s r_{s-1} \\ &= O_P(H^{-1}n^{-1} + (H^{-3}n^{-1} + H^{-2}n^{-1}n^{-1})^{1/2} + (H^{-3} + H^{-2}n^{-1})^{1/2}) = O_P(H^{-3/2}). \end{aligned}$$

For $IE_{3,2}$, we have

$$IE_{3,2} = O_p(p_f H^{-1}n^{-1} + (n \cdot n^{-1}H^{-3}n^{-1})^{1/2} + (nH^{-3})^{1/2}) = O_P(p_f H^{-1}n^{-1} + H^{-3/2}n^{1/2})$$
$$= O_P(H^{-3/2}n^{1/2}),$$

which gives $IE_3 = O_P(H^{-3/2}n^{1/2})$. Hence combining the rates for IE_1 to IE_3, we have

$$K_3 = \begin{cases} O_P(0 + 0 + H^{-3/2}n^{1/2}) = O_P(H^{-3/2}n^{1/2}), & k_H(h) = k((h-1)/H); \\ O_P(0 + H^{-2}n + H^{-3/2}n^{1/2}) = O_P(H^{-2}n), & k_H(h) = k(h/H). \end{cases} \quad \square$$

Proposition 4. *Let all the assumptions in Theorem 4 hold. Then with $p/n \to c \geq 0$,*

$$K_{21} + K_{22} = O_P(H^{-1/2} + n^{-1/4} + p_f^{1/2}n^{-1/2} + H^{-1}n^{1/4} + J^{1/2}n^{-1/2}),$$

where $p_f = 1$ when there are no factors, or $p_f = p$ when there are pervasive factors in the log-price processes.

Proof of Proposition 4. Using the notations used in the proof of Proposition 3, dropping superscripts ij in r_s^{ij}, we can decompose

$$\begin{aligned} K_{21} &= \sum_{s=2}^{n-2J+1} \mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}_{v_s}^{*(J)} - \widetilde{\mathbf{X}}_{v_{s-1}}^{*(J)})\bigg(r_s - r_{s-1} + 2\sum_{h=1}^{s-2} k_H(h)(r_{s-h} - r_{s-h-1})\bigg) \\ &= \sum_{s=2}^{n-2J+1} \mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}_{v_s}^{*(J)} - \widetilde{\mathbf{X}}_{v_{s-1}}^{*(J)})\bigg((r_s + r_{s-1}) + 2(k_H(1) - 1)r_{s-1} \\ &\qquad + 2\sum_{h=2}^{s-2}\big(k_H(h) - k_H(h-1)\big)r_{s-h}\bigg). \end{aligned}$$

Similarly,

$$\begin{aligned} K_{22} &= \sum_{s=2}^{n-2J+1} (r_s - r_{s-1})\bigg(\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}_{v_s}^{*(J)} - \widetilde{\mathbf{X}}_{v_{s-1}}^{*(J)}) + 2\sum_{h=1}^{s-2} k_H(h)\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}_{v_{s-h}}^{*(J)} - \widetilde{\mathbf{X}}_{v_{s-h-1}}^{*(J)})\bigg) \\ &= \sum_{s=2}^{n-2J+1} (r_s - r_{s-1})\bigg(\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}_{v_s}^{*(J)} + \widetilde{\mathbf{X}}_{v_{s-1}}^{*(J)}) + 2(k_H(1) - 1)\mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{X}}_{v_{s-1}} \\ &\qquad + \bigg\{2\sum_{h=2}^{s-2}\big(k_H(h) - k_H(h-1)\big)\mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{X}}_{v_{s-h}}^{*(J)} - 2k_H(s-2)\mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{X}}_{v_1}^{*(J)}\bigg\}\bigg). \end{aligned}$$

Hence we have $K_{21} + K_{22} = \sum_{i=1}^{6} CE_i$, where

$$
\begin{aligned}
CE_1 &= 2 \sum_{s=2}^{n-2J+1} \mathbf{p}_{ij}^{\mathrm{T}} (\widetilde{\mathbf{X}}_{v_s}^{*(J)} r_s - \widetilde{\mathbf{X}}_{v_{s-1}}^{*(J)} r_{s-1}), \\
CE_2 &= 2(k_H(1) - 1) \sum_{s=3}^{n-2J+1} \mathbf{p}_{ij}^{\mathrm{T}} (\widetilde{\mathbf{X}}_{v_s}^{*(J)} - \widetilde{\mathbf{X}}_{v_{s-1}}^{*(J)}) r_{s-1}, \\
CE_3 &= 2(k_H(1) - 1) \sum_{s=3}^{n-2J+1} \mathbf{p}_{ij}^{\mathrm{T}} \widetilde{\mathbf{X}}_{v_{s-1}}^{*(J)} (r_s - r_{s-1}), \\
CE_4 &= 2 \sum_{s=4}^{n-2J+1} \mathbf{p}_{ij}^{\mathrm{T}} (\widetilde{\mathbf{X}}_{v_s}^{*(J)} - \widetilde{\mathbf{X}}_{v_{s-1}}^{*(J)}) \sum_{h=2}^{s-2} k'(\xi_h) H^{-1} r_{s-h}, \\
CE_5 &= 2 \sum_{s=4}^{n-2J+1} (r_s - r_{s-1}) \sum_{h=2}^{s-2} k'(\xi_h) H^{-1} \mathbf{p}_{ij}^{\mathrm{T}} \widetilde{\mathbf{X}}_{v_{s-h}}^{*(J)}, \\
CE_6 &= -2 \sum_{s=3}^{n-2J+1} (r_s - r_{s-1}) k_H(s-2) \mathbf{p}_{ij}^{\mathrm{T}} \widetilde{\mathbf{X}}_{v_1}^{*(J)}.
\end{aligned}
$$

Using $r_1 = r_{n-2J+1} = 0$, we have

$$CE_1 = 0.$$

For $k_H(h) = k((h-1)/H)$, $CE_2 = CE_3 = 0$. Otherwise, using the notations in the proof of Proposition 2 and dropping the superscripts ij in $z_{d,s}^{ij}, z_{v,s}^{ij}$ and e_s^{ij} when no ambiguity arises,

$$
\begin{aligned}
CE_2 &= 2k''(\xi_1')\xi_1 H^{-1} \sum_{s=3}^{n-2J+1} (z_{d,s} + z_{v,s} + e_s) r_{s-1}, \quad 0 \le \xi_1' \le \xi_1 \le H^{-1} \\
&= O_P(H^{-2} \cdot (p_f^{1/2} n^{-1/2} + n^{1/2} \cdot n^{-1/2} + n^{-1/4})) = O_p(H^{-2}).
\end{aligned}
$$

where we used Assumption (D1), (V1), (E3) and Lemma 4. Similarly,

$$
\begin{aligned}
CE_3 &= 2k''(\xi_1')\xi_1 H^{-1} \left(\sum_{s=3}^{n-2J+1} \mathbf{p}_{ij}^{\mathrm{T}} (\widetilde{\mathbf{X}}_{v_{s-1}}^{*(J)} - \widetilde{\mathbf{X}}_{v_s}^{*(J)}) r_s - \mathbf{p}_{ij}^{\mathrm{T}} \widetilde{\mathbf{X}}_{v_2}^{*(J)} r_2 \right) \\
&= O_P(H^{-2} + H^{-2}(J^{1/2} n^{-1/2} + n^{-1/4})) = O_P(H^{-2}).
\end{aligned}
$$

We also have

$$CE_4 = 2\sum_{s=4}^{n-2J+1}(z_{d,s}+z_{v,s}+e_s)\sum_{h=2}^{s-2}k'(\xi_h)H^{-1}r_{s-h} = CE_{4,d}+CE_{4,v}+CE_{4,e},$$

where

$$\begin{aligned}
CE_{4,d} &= O_P(p_f n^{-1} + p_f^{1/2}H^{-1/2}n^{-1/2} + p_f^{1/2}H^{-1/2}) = O_P(p_f^{1/2}H^{-1/2}),\\
CE_{4,v} &= O_P(n\cdot n^{-1}\cdot(H^{-3}+H^{-2}n^{-1}+H^{-2}+H^{-1}n^{-1}+n^{-1})^{1/2})^{1/2}\\
&= O_P(H^{-1/2}+n^{-1/4}),\\
CE_{4,e} &= O_P(H^{-1/2}n^{-1/4}).
\end{aligned}$$

Hence

$$CE_4 = O_P(H^{-1/2}+n^{-1/4}).$$

For CE_5, we have

$$\begin{aligned}
CE_5 &= 2\sum_{s=4}^{n-2J+1} r_s \sum_{h=2}^{s-2}k'(\xi_h)H^{-1}\mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{X}}_{v_{s-h}}^{*(J)} - 2\sum_{s=3}^{n-2J} r_s\sum_{h=2}^{s-1}k'(\xi_h)H^{-1}\mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{X}}_{v_{s-h+1}}^{*(J)}\\
&= -2r_3k'(\xi_2)H^{-1}\mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{X}}_{v_2}^{*(J)}\\
&\quad + 2\sum_{s=4}^{n-2J} r_s\left(\sum_{h=2}^{s-2}k'(\xi_h)H^{-1}\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}_{v_{s-h}}^{*(J)} - \widetilde{\mathbf{X}}_{v_{s-h+1}}^{*(J)}) - k'(\xi_{s-1})H^{-1}\mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{X}}_{v_2}^{*(J)}\right)\\
&= O_P(H^{-1}(J^{1/2}n^{-1/2}+n^{-1/4}) + (n\cdot p_f n^{-2} + n\cdot H^{-1}n^{-1} + n\cdot H^{-2}n^{-1/2})^{1/2}\\
&\quad + H^{-1/2}(J^{1/2}n^{-1/2}+n^{-1/4}))\\
&= O_P(p_f^{1/2}n^{-1/2} + H^{-1/2} + H^{-1}n^{1/4}).
\end{aligned}$$

Finally,

$$\begin{aligned}
CE_6 &= \left(-2\sum_{s=3}^{n-2J} r_s k_H(s-2) + 2\sum_{s=2}^{n-2J} r_s k_H(s-1)\right)\mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{X}}_{v_1}^{*(J)}\\
&= 2\mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\mathbf{X}}_{v_1}^{*(J)}\left(\sum_{s=3}^{n-2J} r_s k'(\xi_{s-1})H^{-1} + r_2 k_H(1)\right)\\
&= O_P((J^{1/2}n^{-1/2}+n^{-1/4})(H^{-1/2}+1)) = O_P(J^{1/2}n^{-1/2}+n^{-1/4}).
\end{aligned}$$

Hence

$$K_{21} + K_{22} = O_P(H^{-1/2} + n^{-1/4} + p_f^{1/2}n^{-1/2} + H^{-1}n^{1/4} + J^{1/2}n^{-1/2}). \ \square$$

To finish the proof of Theorem 2, we can combine the rates from Proposition 1 to 4 to arrive at

$$\max_{\substack{i=1,\dots,p \\ j=1,\dots,L}} \left| \frac{\mathbf{p}_{ij}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)^K \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} - 1 \right|$$
$$= O_P(J^{-1} + J^{1/2}n^{-1/2} + p_f^{1/2}n^{-1/4} + p_f^{1/2}H^{1/2}n^{-1/2} + H^{-3/2}n^{1/2} + H^{-1/2} + H^{-1}n^{1/4}).$$

The optimized rate is achieved using $H \asymp n^{1/2}$, with the smallest order of J being $J \asymp p_f^{-1/2}n^{1/4}$, and rate of convergence at $p_f^{1/2}n^{-1/4}$. Hence when there are pervasive factors, we need $p = o(n^{1/2})$ for guaranteed convergence. Also, we have

$$\max_{\substack{i=1,\dots,p \\ j=1,\dots,L}} \left| \frac{\mathbf{p}_{ij}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)^{KP} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} - 1 \right|$$
$$= O_P(J^{-1} + J^{1/2}n^{-1/2} + p_f^{1/2}n^{-1/4} + p_f^{1/2}H^{1/2}n^{-1/2} + H^{-2}n + H^{-1/2} + H^{-1}n^{1/4}).$$

The optimized rate is achieved using $H \asymp p_f^{-1/5}n^{3/5}$, with the smallest order of J being $J \asymp p_f^{-2/5}n^{1/5}$, and rate of convergence at $p_f^{2/5}n^{-1/5}$. Hence when there are pervasive factors, we need $p = o(n^{1/2})$ for guaranteed convergence.

For $\hat{\mathbf{\Sigma}}(0,1)$, we use the same argument at the end of the proof of Theorem 1 to arrive at

$$\left\| \hat{\mathbf{\Sigma}}(0,1)^K \mathbf{\Sigma}_{\text{Ideal}}(0,1) - \mathbf{I}_p \right\| = O_P(p_f^{1/2}n^{-1/4}),$$

where $H \asymp n^{1/2}$ and $J \asymp p_f^{-1/2}n^{1/4}$. Same treatment apply for $\hat{\mathbf{\Sigma}}(0,1)^{KP}$. This completes the proof of the theorem. $\square$

3.7.3 Proof of the Theorem 3

We first present the proof for $\widetilde{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)^{PP} = \mathbf{P}(\widetilde{\mathbf{Y}})_j$. Let

$$\mathbf{P}(\widetilde{\mathbf{Y}})_j = \frac{1}{\psi Q} \sum_{s=1}^{n-Q+1} \bar{\widetilde{\mathbf{Y}}}_s \bar{\widetilde{\mathbf{Y}}}_s^{\mathrm{T}}, \text{ where } \bar{\widetilde{\mathbf{Y}}}_s = \sum_{l=1}^{Q-1} g\left(\frac{l}{Q}\right)(\widetilde{\mathbf{Y}}_{s+l} - \widetilde{\mathbf{Y}}_{s+l-1}).$$

Then we have

$$\begin{aligned}
\mathbf{P}(\widetilde{\mathbf{Y}})_j =&\mathbf{P}(\widetilde{\mathbf{X}})_j + \mathbf{P}(\widetilde{\mathbf{X}},\mathbf{E})_j + \mathbf{P}(\mathbf{E},\widetilde{\mathbf{X}})_j + \mathbf{P}(\mathbf{E})_j, \text{ where}\\
\mathbf{P}(\widetilde{\mathbf{X}})_j =&\frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\bar{\widetilde{\mathbf{X}}}_{v_s}\bar{\widetilde{\mathbf{X}}}_{v_s}^{\mathrm{T}}, \text{ with } \bar{\widetilde{\mathbf{X}}}_s = \sum_{l=1}^{Q-1} g\Big(\frac{l}{Q}\Big)(\widetilde{\mathbf{X}}_{v_{s+l}} - \widetilde{\mathbf{X}}_{v_{s+l-1}}),\\
\mathbf{P}(\mathbf{E})_j =&\frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\bar{\mathbf{E}}(s)\bar{\mathbf{E}}(s)^{\mathrm{T}}, \text{ with } \bar{\mathbf{E}}_s = \sum_{l=1}^{Q-1} g\Big(\frac{l}{Q}\Big)(\mathbf{E}(s+l) - \mathbf{E}(s+l-1)),\\
\mathbf{P}(\widetilde{\mathbf{X}},\mathbf{E})_j =& \frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\bar{\widetilde{\mathbf{X}}}_{v_s}\bar{\mathbf{E}}(s)^{\mathrm{T}},\ \mathbf{P}(\mathbf{E},\widetilde{\mathbf{X}}) = \mathbf{P}(\widetilde{\mathbf{X}},\mathbf{E})^{\mathrm{T}}.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{P}(\widetilde{\mathbf{Y}})_j\mathbf{p}_{ij} &= \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{P}(\widetilde{\mathbf{X}})_j\mathbf{p}_{ij} + 2\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{P}(\widetilde{\mathbf{X}},\mathbf{E})_j\mathbf{p}_{ij} + \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{P}(\mathbf{E})_j\mathbf{p}_{ij}\\
&= I_1^{PP} + 2I_2^{PP} + I_3^{PP}. \qquad (3.14)
\end{aligned}$$

By the definition of ψ, as $Q\to\infty$, standard rate of convergence of Riemann integral implies that

$$\Big|\frac{1}{\psi Q}\sum_{l=1}^{Q-1} g\Big(\frac{l}{Q}\Big)^2 - 1\Big| = O(Q^{-1}). \qquad (3.15)$$

We further decompose

$$\begin{aligned}
I_1^{PP} =&\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{P}(\widetilde{\mathbf{X}})_j\mathbf{p}_{ij}\\
=&\mathbf{p}_{ij}^{\mathrm{T}}\Big(\frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\Big(\sum_{l=0}^{Q-1} g\Big(\frac{l}{Q}\Big)(\widetilde{\mathbf{X}}_{v_{s+l}} - \widetilde{\mathbf{X}}_{v_{s+l-1}})\sum_{l'=0}^{Q-1} g\Big(\frac{l'}{Q}\Big)(\widetilde{\mathbf{X}}_{v_{s+l'}} - \widetilde{\mathbf{X}}_{v_{s+l'-1}})^{\mathrm{T}}\Big)\Big)\mathbf{p}_{ij}\\
=&\frac{1}{\psi Q}\sum_{l=1}^{Q-1} g^2\Big(\frac{l}{Q}\Big)\mathbf{p}_{ij}^{\mathrm{T}}\sum_{s=1}^{n-Q+1}(\widetilde{\mathbf{X}}_{v_{s+l}} - \widetilde{\mathbf{X}}_{v_{s+l-1}})(\widetilde{\mathbf{X}}_{v_{s+l}} - \widetilde{\mathbf{X}}_{v_{s+l-1}})^T\mathbf{p}_{ij}\\
&+\frac{1}{\psi Q}\sum_{l\neq l'} g\Big(\frac{l}{Q}\Big)g\Big(\frac{l'}{Q}\Big)\mathbf{p}_{ij}^{\mathrm{T}}\sum_{s=1}^{n-Q+1}(\widetilde{\mathbf{X}}_{v_{s+l}} - \widetilde{\mathbf{X}}_{v_{s+l-1}})(\widetilde{\mathbf{X}}_{v_{s+l'}} - \widetilde{\mathbf{X}}_{v_{s+l'-1}})^T\mathbf{p}_{ij},\\
=&I_{1,1}^{PP} + I_{1,2}^{PP}.
\end{aligned}$$

Then by Lemma 3 and (3.15),

$$\Big|\frac{I_{1,1}^{PP}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} - 1\Big| = O_P(p_f^{1/2}n^{-1/2} + Q^{-1}). \qquad (3.16)$$

For $I_{1,2}^{PP}$,

$$
\begin{aligned}
I_{1,2}^{PP} &= 2(\psi Q)^{-1} \sum_{s=1}^{n-Q+1} \sum_{l=2}^{Q-1} g\left(\frac{l}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (\widetilde{\mathbf{X}}_{v_{s+l}} - \widetilde{\mathbf{X}}_{v_{s+l-1}}) \sum_{l'=1}^{l-1} g\left(\frac{l'}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (\widetilde{\mathbf{X}}_{v_{s+l'}} - \widetilde{\mathbf{X}}_{v_{s+l'-1}}) \\
&= 2(\psi Q)^{-1} \sum_{s=1}^{n-Q+1} \sum_{l=2}^{Q-1} g\left(\frac{l}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (\mathbf{X}_{v_{s+l}} - \mathbf{X}_{v_{s+l-1}} + e(\mathbf{J}_{v_{s+l}}) - e(\mathbf{J}_{v_{s+l-1}})) \\
&\qquad \cdot \sum_{l'=1}^{l-1} g\left(\frac{l'}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (\mathbf{X}_{v_{s+l'}} - \mathbf{X}_{v_{s+l'-1}} + e(\mathbf{J}_{v_{s+l'}}) - e(\mathbf{J}_{v_{s+l'-1}})) \\
&= 2(\psi Q)^{-1} (P_1 + P_2 + P_3 + P_4),
\end{aligned}
$$

where

$$
\begin{aligned}
P_1 &= \sum_{s=1}^{n-Q+1} \sum_{l=2}^{Q-1} g\left(\frac{l}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (\mathbf{X}_{v_{s+l}} - \mathbf{X}_{v_{s+l-1}}) \sum_{l'=1}^{l-1} g\left(\frac{l'}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (\mathbf{X}_{v_{s+l'}} - \mathbf{X}_{v_{s+l'-1}}), \\
P_2 &= \sum_{s=1}^{n-Q+1} \sum_{l=2}^{Q-1} g\left(\frac{l}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (\mathbf{X}_{v_{s+l}} - \mathbf{X}_{v_{s+l-1}}) \sum_{l'=1}^{l-1} g\left(\frac{l'}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (e(\mathbf{J}_{v_{s+l'}}) - e(\mathbf{J}_{v_{s+l'-1}})), \\
P_3 &= \sum_{s=1}^{n-Q+1} \sum_{l=2}^{Q-1} g\left(\frac{l}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (e(\mathbf{J}_{v_{s+l}}) - e(\mathbf{J}_{v_{s+l-1}})) \sum_{l'=1}^{l-1} g\left(\frac{l'}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (\mathbf{X}_{v_{s+l'}} - \mathbf{X}_{v_{s+l'-1}}), \\
P_4 &= \sum_{s=1}^{n-Q+1} \sum_{l=2}^{Q-1} g\left(\frac{l}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (e(\mathbf{J}_{v_{s+l}}) - e(\mathbf{J}_{v_{s+l-1}})) \sum_{l'=1}^{l-1} g\left(\frac{l'}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} (e(\mathbf{J}_{v_{s+l'}}) - e(\mathbf{J}_{v_{s+l'-1}})).
\end{aligned}
$$

We focus on P_1 first. Defining

$$
A_s(l) = g\left(\frac{l}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{A}(v_{s+l-1}, v_{s+l}) \mathbf{Z}_{d,s+l}^{j}, \quad V_s(l) = g\left(\frac{l}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(v_{s+l-1}, v_{s+l})^{1/2} \mathbf{Z}_{v,s+l}^{j},
$$

we can decompose $P_1 = P_{a,a} + P_{a,v} + P_{v,a} + P_{v,v}$, where

$$
P_{a,v} = \sum_{s=1}^{n-Q+1} \sum_{l=2}^{Q-1} A_s(l) \sum_{l'=1}^{l-1} V_s(l'), \quad P_{v,a} = \sum_{s=1}^{n-Q+1} \sum_{l=2}^{Q-1} V_s(l) \sum_{l'=1}^{l-1} A_s(l'),
$$

and the other two terms are defined similarly. Since $E_j(A_s(l)) = 0$, and

$$E_j\left(\sum_{l=2}^{Q-1} A_s(l)\right)^2 = \sum_{l=2}^{Q-1} g^2\left(\frac{l}{Q}\right)\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s+l-1}, v_{s+l})\mathbf{A}(v_{s+l-1}, v_{s+l})^{\mathrm{T}}\mathbf{p}_{ij}$$
$$+ \sum_{l\neq l'} g\left(\frac{l}{Q}\right)g\left(\frac{l'}{Q}\right)E_j(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s+l-1}, v_{s+l})\mathbf{Z}_{d,s+l}^j\mathbf{Z}_{d,s+l'}^j\mathbf{A}(v_{s+l'-1}, v_{s+l'})^{\mathrm{T}}\mathbf{p}_{ij})$$
$$= O(Qp_f n^{-2} + Q^2 p_f n^{-2}) = O(Q^2 p_f n^{-2}),$$

we have

$$P_{a,a} = O_P(n \cdot (p_f^{1/2} Q n^{-1})^2) = O_P(p_f Q^2 n^{-1}).$$

Also, Burkholder's inequality can be applied so that for a generic constant C which can change values from line to line,

$$E_j\left(\sum_{l=2}^{Q-1} V_s(l)\right)^2 \leq C\sum_{l=2}^{Q-1} g^2\left(\frac{l}{Q}\right)\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_{s+l-1}, v_{s+l})\mathbf{p}_{ij} = O(Qp_f n^{-1}),$$

so that

$$P_{a,v} = O_P(n \cdot p_f^{1/2} Q n^{-1} \cdot p_f^{1/2} Q^{1/2} n^{-1/2}) = O_P(p_f Q^{3/2} n^{-1/2}).$$

Since $E(V_s(l)\sum_{l'=1}^{l-1} A_s(l')|\mathcal{F}_{-j} \cup \mathcal{F}_{v_{s+l-1}}^j) = 0$, we can use Burkholder's inequality so that

$$E_j\left(\sum_{s=1}^{n-Q+1} V_s(l)\sum_{l'=1}^{l-1} A_s(l')\right)^2 \leq C\sum_{s=1}^{n-Q+1} E_j^{1/2}(V_s(l)^4)E_j^{1/2}\left(\sum_{l'=1}^{l-1} A_s(l')\right)^4$$
$$= O(n \cdot p_f n^{-1} \cdot Q^2 p_f n^{-2}) = O(p_f^2 Q^2 n^{-2}).$$

where the second inequality used Lemma 2.7 of Bai and Silverstein (1998). It means that

$$P_{v,a} = O_P(Q \cdot p_f Q n^{-1}) = O_P(p_f Q^2 n^{-1}).$$

Also, similar to the arguments above,

$$E_j\left(\sum_{s=1}^{n-Q+1} V_s(l)\sum_{l'=1}^{l-1} V_s(l')\right)^2 = O(n \cdot p_f n^{-1} \cdot Q p_f n^{-1}) = O(p_f^2 Q n^{-1}).$$

Hence

$$P_{v,v} = O_P(p_f Q^{3/2} n^{-1/2}).$$

With these results, we have

$$P_1/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij} = O_P(p_f Q^2 n^{-1} + p_f^{1/2} Q^{3/2} n^{-1/2}).$$

Using Assumption (W1) to (W3) and the rate of jump removal at $n^{-1/4}$ from Fan and Wang (2007), we have

$$\frac{P_2}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}, \frac{P_3}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = O_P(p_f^{1/2} Q^2 n^{-5/4} + Q^{3/2} n^{-3/4}),$$
$$\frac{P_4}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = O_P(Q n^{-1/2}).$$

Hence using the rates for P_1 to P_4 found above, together with (3.16), we have

$$\left| \frac{I_1^{PP}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} - 1 \right| = O_P(p_f^{1/2} Q^{1/2} n^{-1/2} + p_f Q n^{-1} + Q^{-1}). \tag{3.17}$$

We find the rate of I_3^{PP} now. Using $g(0) = g(1) = 0$, we have

$$\begin{aligned}\bar{\mathbf{E}}(s) &= \sum_{l=1}^{Q-1} g\left(\frac{l}{Q}\right)(\mathbf{E}(s+l) - \mathbf{E}(s+l-1)) \\ &= \sum_{l=0}^{Q-1}\left[g\left(\frac{l}{Q}\right) - g\left(\frac{l+1}{Q}\right)\right]\mathbf{E}(s+l) = -Q^{-1}\sum_{l=0}^{Q-1} g'_{e_{l+1}}\mathbf{E}(s+l),\end{aligned}$$

where $l/Q \le e_{l+1} \le (l+1)/Q$. Hence

$$\begin{aligned}&\mathbf{p}_{ij}^{\mathrm{T}}\bar{\mathbf{E}}(s)\bar{\mathbf{E}}(s)^{\mathrm{T}}\mathbf{p}_{ij} = \mathbf{p}_{ij}^{\mathrm{T}}\left(Q^{-1}\sum_{l=0}^{Q-1} g'_{e_{l+1}}\mathbf{E}(s+l)\right)\left(Q^{-1}\sum_{l=0}^{Q-1} g'_{e_{l+1}}\mathbf{E}(s+l)^{\mathrm{T}}\right)\mathbf{p}_{ij} \\ &= Q^{-2}\Big(\sum_{l=0}^{Q-1} g'^2_{e_{l+1}}\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}(s+l)\mathbf{E}(s+l)^{\mathrm{T}}\mathbf{p}_{ij} + \sum_{l\neq l'} g'_{e_{l+1}} g'_{e_{l'+1}}\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}(s+l)\mathbf{E}(s+l')^{\mathrm{T}}\mathbf{p}_{ij}\Big).\end{aligned} \tag{3.18}$$

By Lemma 6 and Lemma 7, we then have

$$\begin{aligned}
&\left|\frac{I_3^{PP}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right| \\
=&O_P\bigg(Q^{-1}Q^{-2}\big(\sum_{l=0}^{Q-1}\sum_{s=1}^{n-Q+1}\frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{E}(s+l)\mathbf{E}(s+l)^{\mathrm{T}})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}+ \\
&\qquad 2\sum_{l'<l}\sum_{s=1}^{n-Q+1}\frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{E}(s+l)\mathbf{E}(s+l')^{\mathrm{T}})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\big)\bigg) \qquad (3.19)\\
=&O_P\big(Q^{-2}(n)+Q^{-3/2}(n^{1/2})\big)=O_P\big(Q^{-2}n\big). \qquad (3.20)
\end{aligned}$$

For $I_2^{PP}/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}$, it can be decomposed as $\sum_{i=1}^{6}L_i/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}$, where

$$\begin{aligned}
L_1=&\frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\sum_{l=0}^{Q-1}g\bigg(\frac{l}{Q}\bigg)\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}_{v_{s+l}}-\mathbf{X}_{v_{s+l-1}})\sum_{l'=0}^{Q-1}(-g_{e'_{l'+1}})Q^{-1}\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s+l'),\\
L_2=&\frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\sum_{l=0}^{Q-1}g\bigg(\frac{l}{Q}\bigg)\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}_{v_{s+l}}-\mathbf{X}_{v_{s+l-1}})\sum_{l'=0}^{Q-1}(-g_{e'_{l'+1}})Q^{-1}\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s+l')-\mathbf{X}_{v_{s+l'}}),\\
L_3=&\frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\sum_{l=0}^{Q-1}g\bigg(\frac{l}{Q}\bigg)\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}_{v_{s+l}}-\mathbf{X}_{v_{s+l-1}})\\
&\qquad\cdot\sum_{l'=0}^{Q-1}(-g_{e'_{l'+1}})Q^{-1}\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}(s+l'))-e(\mathbf{J}_{v_{s+l'}})),\\
L_4=&\frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\sum_{l=0}^{Q-1}g\bigg(\frac{l}{Q}\bigg)\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}_{v_{s+l}})-e(\mathbf{J}_{v_{s+l-1}}))\sum_{l'=0}^{Q-1}(-g_{e_{l'+1}})Q^{-1}\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s+l'),\\
L_5=&\frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\sum_{l=0}^{Q-1}g\bigg(\frac{l}{Q}\bigg)\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}_{v_{s+l}})-e(\mathbf{J}_{v_{s+l-1}}))\\
&\qquad\cdot\sum_{l'=0}^{Q-1}(-g_{e_{l'+1}})Q^{-1}\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s+l')-\mathbf{X}_{v_{s+l'}}),\\
L_6=&\frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\sum_{l=0}^{Q-1}g\bigg(\frac{l}{Q}\bigg)\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}_{v_{s+l}})-e(\mathbf{J}_{v_{s+l-1}}))\\
&\qquad\cdot\sum_{l'=0}^{Q-1}(-g_{e_{l'+1}})Q^{-1}\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}(s+l'))-e(\mathbf{J}_{v_{s+l'}})).
\end{aligned}$$

Using notations as before, we have

$$\begin{aligned}
\sum_{l=0}^{Q-1} g\left(\frac{l}{Q}\right) \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}_{v_{s+l}} - \mathbf{X}_{v_{s+l-1}})}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}} &= \sum_{l=0}^{Q-1} \frac{A_s(l) + V_s(l)}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}} \\
&= O_P(p_f^{1/2} Q n^{-1} + Q^{1/2} n^{-1/2}), \\
\sum_{l=0}^{Q-1} \frac{(-g_{e_{l+1}}) Q^{-1} \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s+l)}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}} &= O_P(Q \cdot Q^{-2} + Q^2 \cdot Q^{-2} n^{-1})^{1/2} = O_P(Q^{-1/2}), \\
\sum_{l=0}^{Q-1} (-g_{e_{l+1}}) Q^{-1} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s+l) - \mathbf{X}_{v_{s+l}})}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}} &= O_P(p_f n^{-1} + Q^{-1} Q^{1/2} n^{-1/2}) \\
&= O_P(p_f^{1/2} n^{-1} + Q^{-1/2} n^{-1/2}).
\end{aligned}$$

With these, together with Assumption (W1) to (W3) and the rate of $n^{-1/4}$ in Fan and Wang (2007) for jump removal , we have

$$\begin{aligned}
L_3/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij} &= O_P((p_f^{1/2} Q n + Q^{1/2} n^{-1/2}) \cdot Q^{-1} n^{-1/4}) \\
&= O_P(p_f^{1/2} n^{-5/4} + Q^{-1/2} n^{-3/4}), \\
L_4/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij} &= O_P(Q^{-1} \cdot Q^{-1/2} \cdot Q n^{-1/4}) = O_P(Q^{-1/2} n^{-1/4}), \\
L_5/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij} &= O_P(Q^{-1} Q n^{-1/4} (p_f^{1/2} n^{-1} + Q^{-1/2} n^{-1/2})) \\
&= O_P(p_f^{1/2} n^{-5/4} + Q^{-1/2} n^{-3/4}), \\
L_6/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij} &= O_P(Q^{-1} \cdot Q n^{-1/4} \cdot n^{-1/4}) = O_P(n^{-1/2}).
\end{aligned}$$

For L_1, consider $L_1 = \sum_{l=0}^{Q-1} L_{1,l'}$, where

$$L_{1,l'} = \frac{1}{\psi Q} \sum_{s=1}^{n-Q+1} \sum_{l=0}^{Q-1} g\left(\frac{l}{Q}\right) \mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}_{v_{s+l}} - \mathbf{X}_{v_{s+l-1}})(-g_{e'_{l'+1}}) Q^{-1} \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s+l').$$

Then using Assumption (E3),

$$\begin{aligned}
E_j(L_{1,l'}^2/(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^2) &= O(Q^{-2} \cdot n \cdot (p_f^{1/2} Q n^{-1} + Q^{1/2} n^{-1/2})^2 Q^{-2}) \\
&= O(p_f Q^{-2} n^{-1} + Q^{-3}).
\end{aligned}$$

Hence

$$L_1/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij} = O_P(p_f^{1/2} n^{-1/2} + Q^{-1/2}).$$

We also have

$$\begin{aligned} L_2/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij} &= O_P(Q^{-1}n(p_fQn^{-1}+Q^{1/2}n^{-1/2})(p_f^{1/2}n^{-1}+Q^{-1/2}n^{-1/2})) \\ &= O_P(p_fn^{-1}+Q^{-1}). \end{aligned}$$

Hence we have

$$\left|\frac{I_2^{PP}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right| = O_P(p_f^{1/2}n^{-1/2}+Q^{-1/2}). \tag{3.21}$$

The above results imply that

$$\left|\frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{P}(\widetilde{\mathbf{Y}})_j\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}-1\right| = O_P(Q^{-1/2}+Q^{-2}n+p_f^{1/2}Q^{1/2}n^{-1/2}+p_fQn^{-1}).$$

When there are no factors, $p_f = 1$, and so the best rate is to balance $Q^{-2}n$ and $Q^{1/2}n^{-1/2}$, resulting in $Q \asymp n^{3/5}$ and a rate of convergence at $n^{-1/5}$. When there are pervasive factors such that $p_f = p$, then we need $Q \asymp n^{3/5}p^{-1/5}$, and rate of convergence at $n^{-1/5}p^{2/5}$. Hence we need $p = o(n^{1/2})$ for guaranteed convergence in this case.

This completes the proof for the positive semi-definite version, since the above rate does not depend on both the indices i and j.

Consider now NER-PRVM which is $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1},\tau_j)^P$. Using the same decomposition as $\mathbf{P}(\widetilde{\mathbf{Y}})_j$, we have

$$\begin{aligned} \mathbf{p}_{ij}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1},\tau_j)^P\mathbf{p}_{ij} &= \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{P}(\widetilde{\mathbf{Y}})_j\mathbf{p}_{ij} - \varsigma(\psi Q)^{-1}\sum_{s=1}^{n-Q+1}\mathbf{p}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\eta}}^{(j)}\mathbf{p}_{ij}, \\ &= I_1^P + 2I_2^P + I_3^P, \end{aligned}$$

where I_1^P, and I_2^P are exactly the same as I_1^{PP} and I_2^{PP} from the decomposition of $\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{P}(\widetilde{\mathbf{Y}})\mathbf{p}_{ij}$ at the beginning of the proof, and

$$I_3^P = \frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\left[\mathbf{p}_{ij}^{\mathrm{T}}\bar{\mathbf{E}}(s)\bar{\mathbf{E}}(s)^{\mathrm{T}}\mathbf{p}_{ij} - \varsigma\cdot\mathbf{p}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\eta}}^{(j)}\mathbf{p}_{ij}\right].$$

From (3.17) and (3.21), we have

$$\left|\frac{I_1^P}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}-1\right| = O_P(p_f^{1/2}Q^{1/2}n^{-1/2}+p_fQn^{-1}+Q^{-1}),$$
$$\left|\frac{I_2^P}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right| = O_P(p_f^{1/2}n^{-1/2}+Q^{-1/2}).$$

We can decompose $I_3^P = J_1 + J_2$, where

$$J_1 = \frac{1}{\psi Q}\sum_{s=1}^{n-Q+1}\left[\mathbf{p}_{ij}^{\mathrm{T}}\bar{\mathbf{E}}(s)\bar{\mathbf{E}}(s)^{\mathrm{T}}\mathbf{p}_{ij}-\varsigma\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^j\mathbf{p}_{ij}\right],$$
$$J_2 = \frac{\varsigma}{\psi Q}\sum_{s=1}^{n-Q+1}(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^j\mathbf{p}_{ij}-\mathbf{p}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\eta}}^{(j)}\mathbf{p}_{ij}).$$

Since $\bar{\mathbf{E}}(s)=\sum_{l=1}^{Q-1}[g(l/Q)-g((l+1)/Q)]E(s+l)=Q^{-1}\sum_{l=1}^{Q-1}(-g'_{e_{l+1}})\mathbf{E}(s+l)$, we have

$$J_1 = J_{1,1}+J_{1,2},\ \text{where}$$
$$J_{1,1} = \frac{1}{\psi Q}\sum_{l=1}^{Q-1}\left[g\left(\frac{l}{Q}\right)-g\left(\frac{l+1}{Q}\right)\right]^2\sum_{s=1}^{n-Q+1}(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}(s+l)\mathbf{E}(s+l)^{\mathrm{T}}\mathbf{p}_{ij}-\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^j\mathbf{p}_{ij}),$$
$$J_{1,2} = \psi^{-1}Q^{-3}\sum_{l\neq l'}g'_{e_{l+1}}g'_{e_{l'+1}}\sum_{s=1}^{n-Q+1}\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}(s+l)\mathbf{E}(s+l')^{\mathrm{T}}\mathbf{p}_{ij}.$$

From (3.19), we know that

$$J_{1,2}/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij} = O_P(Q^{-3/2}n^{1/2}). \tag{3.22}$$

To find the rate of $J_{1,1}$, for $l<Q$, it suffice to consider the rate of

$$\sum_{s=1}^{n-Q+1}\frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}(s+l)\mathbf{E}(s+l)^{\mathrm{T}}\mathbf{p}_{ij}-\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^j\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = E'_{1,1}+E'_{1,2}+E_2+E_3+2\sum_{k=4}^{6}E_k,$$

where E_2 to E_6 are essentially the same as those defined in the proof of Lemma 6

(only with s replaced by $s+l$ and n by $n-Q+1$), and

$$E'_{1,1} = \sum_{s=1}^{n-Q+1} \frac{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s+l)\boldsymbol{\epsilon}(s+l)^{\mathrm{T}} \mathbf{p}_{ij} - \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\epsilon,s+l}^{j} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}},$$
$$E'_{1,2} = \sum_{s=1}^{n-Q+1} \frac{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\epsilon,s+l}^{j} \mathbf{p}_{ij} - \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}_{\epsilon,s}^{j} \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}}.$$

Under Assumption (E3), using Lemma 2.7 of Bai and Silverstein (1998), we have $E_j(E'_{1,1}) = 0$, and

$$E_j\big((E'_{1,1})^2 | \boldsymbol{\Sigma}_{\epsilon,u}, u \in [0,1]\big) = O(n + n^2 \cdot n^{-1}) = O(n).$$

Hence

$$E'_{1,1} = O_P(n^{1/2}).$$

Also, by Assumption (E1) on the smoothness of $\{\boldsymbol{\Sigma}_{\epsilon,s}\}_s$, we have

$$E'_{1,2} = O_P(|l|).$$

From the proof of Lemma 6, we have

$$E_2 = O_P(1),\ E_3 = O_P(n^{-1/2}).$$

Using Assumption (W1) to (W3) and the jump removal rate of $n^{-1/4}$ from Fan and Wang (2007),

$$E_5 = O_P(n^{-1/4}),\ E_6 = O_P(n^{-3/4}).$$

For E_4, using Assumption (E3),

$$E_4 = O_P(n \cdot (n^{-1/2})^2 + n^2 \cdot n^{-1} \cdot (n^{-1/2})^2) = O_P(1).$$

From these rates, we have

$$\frac{J_{1,1}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} = O_P\bigg(\frac{1}{\psi Q} \sum_{l=1}^{Q-1} (g'_{e_{l+1}})^2 Q^{-2}(n^{1/2} + l)\bigg) = O_P(Q^{-1} + Q^{-2} n^{1/2}).$$

Hence

$$J_1/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij} = O_P(Q^{-1} + Q^{-3/2}n^{1/2}) = O_P(Q^{-3/2}n^{1/2}).$$

For J_2, consider

$$\begin{aligned}
&\frac{\mathbf{p}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\eta}}^{(j)}\mathbf{p}_{ij} - \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = \sum_{i=1}^{5} R_i, \text{ where} \\
R_1 &= \frac{1}{2n}\sum_{s=2}^{n}\frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}(s)\mathbf{E}(s)^{\mathrm{T}}\mathbf{p}_{ij} - \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} - \frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij}}{2n\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}, \\
R_2 &= \frac{1}{2n}\sum_{s=2}^{n}\frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}(s-1)\mathbf{E}(s-1)^{\mathrm{T}}\mathbf{p}_{ij} - \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} - \frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij}}{2n\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}, \\
R_3 &= \frac{-1}{n}\sum_{s=2}^{n}\frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{E}(s)\mathbf{E}(s-1)^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}, \\
R_4 &= \frac{1}{2n}\sum_{s=2}^{n}\frac{\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}_{v_s} - \widetilde{\mathbf{X}}_{v_{s-1}})(\widetilde{\mathbf{X}}_{v_s} - \widetilde{\mathbf{X}}_{v_{s-1}})^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}, \\
R_5 &= \frac{1}{n}\sum_{s=2}^{n}\frac{\mathbf{p}_{ij}^{\mathrm{T}}(\widetilde{\mathbf{X}}_{v_s} - \widetilde{\mathbf{X}}_{v_{s-1}})(\mathbf{E}(s) - \mathbf{E}(s-1))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}.
\end{aligned}$$

From the proof of the rate for J_1 before, and Assumption (E1), we get

$$R_1, R_2 = O_P(n^{-1}\cdot n^{1/2} + n^{-1}) = O_P(n^{-1/2}).$$

From Lemma 7, we have

$$R_3 = O_P(n^{-1}\cdot n^{1/2}) = O_P(n^{-1/2}).$$

From the result of Lemma 3, we immediately have

$$R_4 = O_P(n^{-1}(n^{-1/2} + p_f^{1/2}n^{-1/2}) + n^{-1}) = O_P(n^{-1}).$$

Substituting $K_m = 1$ in (3.6) and incorporating Assumption (A4), we have

$$R_5 = O_P(n^{-1}\cdot 1) = O_P(n^{-1}).$$

These rates imply that, with $\varsigma = O(Q^{-1})$,

$$J_2 = O_P(Q^{-2} \cdot n \cdot n^{-1/2}) = O_P(Q^{-2} n^{1/2}).$$

Finally we have

$$\left| \frac{I_3^P}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} \right| = O_P(Q^{-3/2} n^{1/2} + Q^{-2} n^{1/2}) = O_P(Q^{-3/2} n^{1/2}).$$

Hence

$$\begin{aligned}
&\max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \frac{\mathbf{p}_{ij}^{\mathrm{T}} \widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^P \mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} - 1 \right| \\
&\leq \max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \frac{I_1^P}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} - 1 \right| + 2 \max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \frac{I_2^P}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} \right| \\
&\qquad + \max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \frac{I_3^P}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} \right| \\
&= O_P(p_f^{1/2} Q^{1/2} n^{-1/2} + Q^{-1/2} + Q^{-3/2} n^{1/2}).
\end{aligned}$$

When there are no factors such that $p_f = 1$, the above rates imply that $Q \asymp n^{1/2}$, with a rate of convergence at $n^{-1/4}$. When there are pervasive factors such that $p_f = p$, the best rate can be achieved with $Q \asymp n^{1/2} p^{-1/4}$, with rate of convergence at $p^{3/8} n^{-1/4}$. So we need $p = o(n^{2/3})$ for guaranteed convergence. This completes the proof for the bias-corrected version.

Finally, using similar arguments at the end of the proof of Theorem 1 in Section 3.7.1, we have

$$\left\| \widehat{\boldsymbol{\Sigma}}(0,1)^P \boldsymbol{\Sigma}_{\mathrm{Ideal}}(0,1)^{-1} - \mathbf{I}_p \right\| = O_P(p_f^{1/2} Q^{1/2} n^{-1/2} + Q^{-1/2} + Q^{-3/2} n^{1/2}),$$

the same as for $\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)^P$. Similar treatments apply for $\widehat{\boldsymbol{\Sigma}}(0,1)^{PP}$. This completes the proof of the theorem. $\square$

3.7.4 Proof of the Theorem 4

We have shown in the proofs of Theorem 1 to 3 for the results using jumps-removed data. To complete the proof, note that

$$\begin{aligned}\Big\| \sum_{0\le t\le 1} (\Delta \mathbf{J}_t \Delta \mathbf{J}_t^{\mathrm{T}} - \Delta \widehat{\mathbf{J}}_t \Delta \widehat{\mathbf{J}}_t^{\mathrm{T}}) \Big\| &\le C \max_{0\le t\le 1} \big\| \Delta \mathbf{J}_t \Delta \mathbf{J}_t^{\mathrm{T}} - \Delta \widehat{\mathbf{J}}_t \Delta \widehat{\mathbf{J}}_t^{\mathrm{T}}) \big\| \\ &\le 2C \max_{0\le t\le 1} \big\| \Delta \mathbf{J}_t - \Delta \widehat{\mathbf{J}}_t \big\| \cdot \big\| \Delta \mathbf{J}_t \big\| + C \max_{0\le t\le 1} \big\| \Delta \mathbf{J}_t - \Delta \widehat{\mathbf{J}}_t \big\|^2 \\ &= O_P(n^{-1/4} L^{-1/4}).\end{aligned}$$

Since L is finite, this completes the proof of the theorem. $\square$

Bibliography

Abreu, M., de Groot, H. L., and Florax, R. J. (2005). Space And Growth: A Survey Of Empirical Evidence And Methods. *Region et Developpement*, 21:13–44.

Ahrens, A. and Bhattacharjee, A. (2015a). Two-step lasso estimation of the spatial weights matrix. *Econometrics*, 3(1):128.

Ahrens, A. and Bhattacharjee, A. (2015b). Two-step lasso estimation of the spatial weights matrix. *Econometrics*, 3(1):128.

Aït-Sahalia, Y., Fan, J., and Xiu, D. (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association*, 105(492):1504–1517.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858.

Anselin, L. (2010). *Spatial Econometrics: Methods and Models*. Studies in Operational Regional Science. Springer, softcover reprint of hardcover 1st ed. 1988 edition.

Anselin, L., Gallo, J. L., and Jayet, H. (2008). Spatial panel econometrics. In Mátyás, L. and Sevestre, P., editors, *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, pages 625–660. Springer Berlin Heidelberg, Berlin, Heidelberg.

Arnold, M., Stahlberg, S., and Wied, D. (2013a). Modeling different kinds of spatial dependence in stock returns. *Empirical Economics*, 44(2):761–774.

Arnold, M., Stahlberg, S., and Wied, D. (2013b). Modeling different kinds of spatial dependence in stock returns. *Empirical Economics*, 44(2):761–774.

Badinger, H. and Egger, P. (2011). Estimation of higher-order spatial autoregressive cross-section models with heteroskedastic disturbances. *Papers in Regional Science*, 90(1):213–235.

Bai, Z. D. and Silverstein, J. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345.

Baltagi, B. (2008). *Econometric analysis of panel data*. John Wiley & Sons.

Baltagi, B. H., Song, S. H., Jung, B. C., and Koh, W. (2007). Testing for serial correlation, spatial autocorrelation and random effects using panel data. *Journal of Econometrics*, 140(1):5–51.

Barndorff-Nielsen, O., Hansen, P. R., Lunde, A., and Shephard, N. (2009). Realized kernels in practice: trades and quotes. *Econometrics Journal*, 12(3):C1–C32.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2011). Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics*, 162(2):149 – 169.

Baumont, C., Ertur, C., and Le Gallo, J. (2003). Spatial convergence clubs and the european regional growth process, 1980–1995. In *European regional growth*, pages 131–158. Springer.

Bhattacharjee, A. and Jensen-Butler, C. (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics*, 43(4):617 – 634.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41 – 55.

Chang, J., Yao, Q., and Zhou, W. (2017). Testing for vector white noise using maximum cross correlations. *Biometrika*, 104(1):1 – 17.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, J. and Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, pages 555–574.

Chen, R. Y. and Mykland, P. A. (2017). Model-free approaches to discern non-stationary microstructure noise and time-varying liquidity in high-frequency data. *Journal of Econometrics*, 200(1):79 – 103.

Christensen, K., Kinnebrock, S., and Podolskij, M. (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of Econometrics*, 159(1):116 – 133.

Cliff, A. and Ord, J. (1973). *Spatial autocorrelation*. Monographs in spatial and environmental systems analysis. Pion.

Corrado, L. and Fingleton, B. (2012). Where is the economics in spatial econometrics? *Journal of Regional Science*, 52(2):210–239.

Dai, C., Lu, K., and Xiu, D. (2017). Knowing factors or factor loadings, or neither? evaluating estimators of large covariance matrices with noisy and asynchronous data. *Forthcoming, Journal of Econometrics*.

Demiguel, V. and Nogales, F. J. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812.

Dou, B., Parrella, M. L., and Yao, Q. (2016). Generalized yulewalker estimation for spatio-temporal models with unknown diagonal coefficients. *Journal of Econometrics*, 194(2):369 – 382.

Elhorst, J. P. (2005). Unconditional maximum likelihood estimation of linear and log-linear dynamic models for spatial panels. *Geographical Analysis*, 37(1):85–106.

Epps, T. W. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 74(366a):291–298.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J., Li, Y., and Yu, K. (2012). Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association*, 107(497):412–428.

Fan, J. and Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association*, 102(480):1349–1362.

Fernandez, V. (2011). Spatial linkages in international financial markets. *Quantitative Finance*, 11(2):237–245.

Foote, C. L. (2007). Space and time in macroeconomic panel data: young workers and state-level unemployment revisited. Working Papers 07-10, Federal Reserve Bank of Boston.

Franzese, R. J. and Hays, J. C. (2007). Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. *Political Analysis*, 15(2):140–164.

Golub, G. H. and van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins University Press, third edition.

Griffin, J. E. and Oomen, R. C. (2011). Covariance measurement in the presence of non-synchronous trading and market microstructure noise. *Journal of Econometrics*, 160(1):58 – 68. Realized Volatility.

Gupta, A. and Robinson, P. M. (2015). Inference on higher-order spatial autoregressive models with increasingly many parameters. *Journal of Econometrics*, 186(1):19–31.

Hallin, M., Lu, Z., Tran, L. T., et al. (2004). Local linear spatial regression. *The Annals of Statistics*, 32(6):2469–2500.

Hayashi, T., Yoshida, N., et al. (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 11(2):359–379.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, second edition. Cambridge Books Online.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, Berkeley, Calif. University of California Press.

Kapoor, M., Kelejian, H. H., and Prucha, I. R. (2007). Panel data models with spatially correlated error components. *Journal of Econometrics*, 140:97–130.

Keller, W. and Shiue, C. H. (2007). The origin of spatial interaction. *Journal of Econometrics*, 140(1):304 – 332. Analysis of spatially dependent data.

Kim, D., Wang, Y., and Zou, J. (2016). Asymptotic theory for large volatility matrix estimation based on high-frequency financial data. *Stochastic Processes and their Applications*, 126(11):3527 – 3577.

Koroglu, M. and Sun, Y. (2016a). Functional-coefficient spatial durbin models with nonparametric spatial weights: An application to economic growth. *Econometrics*, 4(1):6.

Koroglu, M. and Sun, Y. (2016b). Functional-coefficient spatial durbin models with nonparametric spatial weights: An application to economic growth. *Econometrics*, 4(1):6.

Kostov, P. (2013). Choosing the right spatial weighting matrix in a quantile regression model. *ISRN Economics*, 2013.

Lacombe, D. J. (2004). Does econometric methodology matter? an analysis of public policy using spatial econometric techniques. *Geographical Analysis*, 36(2):105–118.

Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Ann. Statist.*, 44(3):928–953.

Lam, C. and Feng, P. (2018). A nonparametric eigenvalue-regularized integrated covariance matrix estimator for asset return data. *Journal of Econometrics*, 206(1):226 – 257.

Lam, C. and Qian, C. (2017). Spatial lag model with time-lagged effects and spatial weight matrix estimation. *Manuscript*.

Lam, C. and Souza, P. C. (2015a). One-step regularized spatial weight matrix and fixed effects estimation with instrumental variables. *Manuscript*.

Lam, C. and Souza, P. C. (2018). Estimation and selection of spatial weight matrix in a spatial lag model. *Journal of Business and Economic Statistics. forthcoming.[Google Scholar]*.

Lam, C. and Souza, P. C. L. (2014). Regularization for spatial panel time series using the adaptive lasso. *LSE STICERD, Econometrics Paper Series*.

Lam, C. and Souza, P. C. L. (2015b). Detection and estimation of block structure in spatial weight matrix. *Econometric Reviews*, pages 1–30.

Lam, C. and Souza, P. C. L. (2015c). One-step regularized spatial weight matrix and fixed effects estimation with instrumental variables. Manuscript.

Lam, C. and Souza, P. C. L. (2016a). Detection and estimation of block structure in spatial weight matrix. *Econometric Reviews*, 35(8-10):1347–1376.

Lam, C. and Souza, P. C. L. (2016b). Spatial lag model estimation with sparse adjustment for spatial weight matrix. Manuscript.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, 40(2):694–726.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060.

Lee, L. and Liu, X. (2010a). Efficient GMM estimation of high order spatial autoregressive models with autoregressive disturbances. *Econometric Theory*, 26:187–230.

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6):1899–1925.

Lee, L.-f. and Liu, X. (2010b). Efficient gmm estimation of high order spatial autoregressive models with autoregressive disturbances. *Econometric Theory*, 26(1):187–230.

Lee, L. F. and Yu, J. (2010). A spatial dynamic panel data model with both time and individual fixed effects. *Econometric Theory*, 26:564–597.

Lee, L.-f. and Yu, J. (2014a). Efficient gmm estimation of spatial dynamic panel data models with fixed effects. *Journal of Econometrics*, 180(2):174–197.

Lee, L. F. and Yu, J. (2014b). Efficient {GMM} estimation of spatial dynamic panel data models with fixed effects. *Journal of Econometrics*, 180(2):174 – 197.

LeSage, J. P. and Pace, R. K. (2008). Spatial econometric modeling of origin-destination flows. *Journal of Regional Science*, 48(5):941–967.

LeSage, J. P. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman and Hall.

Li, K. (2017). Fixed-effects dynamic spatial panel data models and impulse response analysis. *Journal of Econometrics*, 198(1):102–121.

McMillen, D. P., Singell Jr, L. D., and Waddell, G. R. (2007). Spatial competition and the price of college. *Economic Inquiry*, 45(4):817–833.

Robinson, P. (2011). Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics*, 165(1):5 – 19. Moment Restriction-Based Econometric Methods.

Tao, J. (2005). *Spatial econometrics: Models, methods and applications*. PhD thesis, The Ohio State University.

Tao, M., Wang, Y., and Chen, X. (2013). Fast convergence rates in estimating large volatility matrices using high-frequency financial data. *Econometric Theory*, 29(4):838–856.

Tran, L. and Yakowitz, S. (1993). Nearest neighbor estimators for random fields. *Journal of Multivariate Analysis*, 44(1):23 – 46.

Ullah, A. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics: Regression models with an anselin bera i. introduction. In *Handbook of Applied Economic Statistics*, pages 257–259. CRC Press.

van de Geer, S. A. (2002). On hoeffding's inequality for dependent random variables. In Dehling, H., Mikosch, T., and Sørensen, M., editors, *Empirical Process Techniques for Dependent Data*, pages 161–169, Boston, MA. Birkhäuser Boston.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14150–14154.

Wu, W. B. (2011). Asymptotic theory for stationary processes. *Statistics and Its Interface, 0*, pages 1–20.

Xiu, D. (2010). Quasi-maximum likelihood estimation of volatility with high frequency data. *Journal of Econometrics*, 159(1):235 – 250.

Yu, J., de Jong, R., and Lee, L.-f. (2008). Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and t are large. *Journal of Econometrics*, 146(1):118 – 134.

Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach. *Bernoulli*, 12(6):1019–1043.

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33 – 47.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411.

Zhu, X., Pan, R., Li, G., Liu, Y., Wang, H., et al. (2017). Network vector autoregression. *The Annals of Statistics*, 45(3):1096–1123.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.